LOCATING IDEAL CO_2 SEQUESTRATION SITES: AN INTEGRATED ENVIRONMENTAL-BIG DATA ANALYTICS APPROACH

BY

MUHAMMAD SAID ABDALLAH

A Thesis Presented to the

DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

ENVIRONMENTAL SCIENCES

DECEMBER 2019

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN- 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **MUHAMMAD SAID ABDALLAH** under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN ENVIRONMENTAL SCIENCES**.

Dr. Abdullatif Al-Shuhail
Department Chairman

Dr. Mohammad H. Makkawi
(Advisor)

Dr. Salam A. Zummo
Dean of Graduate Studies

Dr. Abdulazeez Abdulraheem
(Member)

31/12/09
Date

Dr. Abdullatif Al-Shuhail
(Member)

To everyone who helped in bringing this work into a reality.

ACKNOWLEDGEMENTS

Acknowledgment is due to the King Fahd University of Petroleum & Minerals for supporting this research.

I wish to express my appreciation to Dr. Mohammad H. Makkawi who served as my major advisor. I also wish to thank other members of my thesis committee Dr. Abdulazeez Abdulraheem and Dr. Abdullatif Al-Shuhail.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAI	:	Absolute Acoustic Impedance
AAPE	:	Average Absolute Percentage Error
ACE	:	Alternating Conditional Expectations
AI	:	Artificial Intelligence
ANFIS	:	Adaptive Neuro-Fuzzy Inference System
ANN	:	Artificial Neural Network
BNN	:	Bayesian Neural Network
BOEM	:	Bureau of Ocean Energy Management
BPNN	:	Back-Propagation Neural Network
CC	:	Correlation Coefficient
CCK	:	Collocated Cokriging
CCS	:	Carbon Capture and Storage
CCTV	:	Closed Circuit Television
CCUS	:	Carbon Capture, Utilization and Storage
CI	:	Coloured Inversion
CNN	:	Convolutional Neural Network
CNN	:	Committee Neural Network
EGR	:	Enhanced Gas Recovery
EOR	:	Enhanced Oil Recovery
FFBP	:	Feed-Forward Back Propagation
FL	:	Fuzzy Logic
GAM	:	Generalized Additive Model
GHG	:	Green House Gas
GR	:	Gamma Ray
GRNN	:	Generalized Regression Neural Network
GUI	:	Graphical User Interface
ICA	:	Imperialist Competitive Algorithm
IOGP	:	Integrated Ocean Drilling Program
LWD	:	Logging While Drilling

MBI	:	Model-Based Inversion
MLFN	:	Multi-Layer Feed Forward Neural Network
MLP	:	Multi-Layer Perceptron
MLP	:	Multi-Linear Perception
MSE	:	Mean Squared Error
NMR	:	Nuclear Magnetic Resonance
NN	:	Neural Network
NNET	:	Neural Network
NN-MAT	:	Neural Network Multi-Attribute Transform
NPHI	:	Neutron Porosity
PNN	:	Probabilistic Neural Network
PVT	:	Pressure, Volume and Temperature
QC	:	Quality Control
$\mathbf{R^2}$	:	Coefficient of Determination
RAI	:	Relative Acoustic Impedance
RBF	:	Radial Basis Function
RHOB	:	Bulk Density
RKB	:	Rotary Kelly Bushing
RMSE	:	Root Mean Squared Error
ROC	:	Receiver Operating Characteristic
SAE	:	Stacked Auto-Encoder
SBLLM	:	Sensitivity Based Linear Learning Method
SMOTE	:	Synthetic Minority Over-Sampling Techniques
SVM	:	Support Vector Machine
TD	:	Total Depth
TVD	:	True Vertical Depth

ABSTRACT

Full Name : MUHAMMAD SAID ABDALLAH

Thesis Title : Locating Ideal CO_2 Sequestration Sites: An Integrated Environmental-Big Data Analytics Approach

Major Field : Environmental Sciences

Date of Degree : December, 2019

The modern industrialization of today's economy has allowed people to enjoy several economic benefits and helped increase their standards of living. However, this industrialization has come with a concerning cost to the Earth's environment. The mass burning of non-renewable fossil fuels releases a great deal of CO_2 into the atmosphere. Research has shown that CO_2 is the main compound that is negatively affecting the stability of Earth's climate. Although there are several natural sources of CO_2 emission to the atmosphere, it is believed that the anthropogenic CO_2 emissions, caused by the excessive exploitation of fossil fuels, is upsetting the natural balance of CO_2. In light of this global environmental issue, researchers have attempted to devise ways to tackle the problem. Several approaches can be adopted to resolve this issue. One such approach is the sequestration of CO_2 in deep geological formations such as depleted oil and gas reservoirs, unmineable underground coal seams and deep non-potable water aquifers.

There are a number of factors that need to be taken into account when attempting to locate a possible CO_2 sequestration site. These factors are akin to the ones that are used to locate a hydrocarbon reservoir for oil and gas exploration purposes. The approach, therefore, we wish to take to locate ideal CO_2 sequestration sites would be akin to locating hydrocarbon prospects.

The research problem can be split into three main sections which can be further subdivided into subsections (data to be used have been put in between parentheses):

- **Locating your Reservoir**: (i) Good Quality Lithologies (log + core); (ii) Favorable Permeability (log + core); (iii) Reservoir Extension (log + seismic); (iv) Net-to-Gross Ratio (log + core).
- **Locating your Seal**: (i) Typical Seal Type Lithologies (log + core); (ii) Low Permeability Layers (log + core); (iii) Seal Extension over Reservoir (log + seismic); (iv) Sufficient Thickness (log + seismic).
- **Locating your Trap**: (i) Presence of Structures (seismic); (ii) Presence of Stratigraphic Traps (seismic); (iii) Presence of Four-Way Dip Closures (seismic); (iv) Presence of Three-Way Dip Closures (seismic).

The research objective is therefore to use the available Big Data analytics and Artificial Intelligence techniques and tools that are available from the Upstream Oil & Gas Industry

to help in the identification of potential CO_2 sequestration reservoirs by the utilization of data such as well log data, core data, and seismic data.
Prediction of formation permeability obtained from core plugs in an exploratory well drilled in the Volve field located in the Norwegian continental shelf via the use of a single layered ANN model is presented. The input data that was fed into the network were three common logs that include the gamma ray (GR) log, the bulk density (RHOB) log, and the neutron porosity (NPHI) log. The optimum single layered ANN model was determined to be one with 18 neurons in the hidden layer, and a log-sigmoidal transfer function. The optimum number of randomly selected data points from the total of 557 data points was found to be 225.

Data obtained from lithology descriptions in logs of the same well were then used in a separate study aiming to classify lithology. The use of ANN in the classification of lithology is important to the determination of suitable CO_2 sequestration sites since the lithology plays a major role in the rock properties (such as permeability and porosity) that would be present. These properties are important in order to permit the movement of CO_2 into the pore space once injected into the formation via injection wells.

Porosity values were also predicted from well log data by use of a single layered ANN. Three input logs were used that were: Gamma Ray (GR), Bulk Density (RHOB), and Neutron Porosity. Log values were correlated with corresponding core porosity values obtained from core analysis of 664 core plugs. The dataset is split into two sets: 70% Training, and 30% Testing. The optimization of the results is achieved by comparing three main statistical parameters that include: the correlation coefficient (CC), the root mean-squared error (RMSE) and the absolute average percentage error (AAPE). The optimum number of neutrons was found to be 18, where the tan-sigmoid transfer function was used in the network. The CC was 0.808, RMSE was 0.055, and AAPE was 5.94. The results illustrate that ANN may be used as a tool in the prediction of porosity in unexplored areas for the sake of characterizing a CO_2 sequestration site. The porosity would play a major factor in the determination of how much CO_2 may be stored in a particular site.

ملخص الرسالة

الاسم الكامل: محمد سعيد عبدالله

عنوان الرسالة: تحديد المواقع المثالية لعزل ثاني أكسيد الكربون: منهج متكامل لتحليل البيانات البيئية الهائلة

التخصص: العلوم البيئية

تاريخ الدرجة العلمية: ديسمبر 2019

لقد أتاح التصنيع الحديث للاقتصاد المعاصر للناس التمتع بالعديد من المزايا الاقتصادية وساعد على زيادة مستويات معيشتهم. بيد أن هذا التصنيع قد أتى أثر بشكل كبيل على بيئة الأرض. حيث يؤدي احتراق الوقود الأحفوري غير المتجدد على نطاق واسع إلى انبعاث كمية كبيرة من ثاني أكسيد الكربون في الغلاف الجوي. وأظهرت الأبحاث أن ثاني أكسيد الكربون هو المركب الرئيسي الذي يؤثر سلبًا على استقرار مناخ الأرض. وعلى الرغم من وجود عدة مصادر طبيعية لانبعاثات ثاني أكسيد الكربون في الغلاف الجوي، يُعتقد أن انبعاثات ثاني أكسيد الكربون البشرية المنشأ الناجمة عن الاستغلال المفرط للوقود الأحفوري تخل بالتوازن الطبيعي لثاني أكسيد الكربون. وفي ضوء هذه المشكلة البيئية العالمية، حاول الباحثون ابتكار سبل لمعالجة تلك المشكلة. ويمكن اعتماد عدة مناهج لحل هذه المشكلة. ومن بين هذا المناهج عزل ثاني أكسيد الكربون في التكوينات الجيولوجية العميقة مثل مستودعات النفط والغاز المستنفدة، وطبقات الفحم الجوفية غير القابلة للاستخراج، ومستودعات المياه الجوفية العميقة غير الصالحة للشرب.

وهناك عدد من العوامل التي ينبغي أخذها في الاعتبار عند محاولة تحديد موقع عزل محتمل لثاني أكسيد الكربون. وهذه العوامل مماثلة لتلك المستخدمة لتحديد مواقع مستودع الهيدروكربونات لأغراض استكشاف النفط والغاز. وبالتالي، فإن المنهج الذي نريد اتباعه لتحديد موقع مثالي لعزل ثاني أكسيد الكربون سيكون أقرب إلى تحديد مواقع الهيدروكربونات.

ويمكن تقسيم مشكلة البحث إلى ثلاثة أقسام رئيسية يتم تقسيمها إلى أقسام فرعية أخرى (وتم وضع البيانات التي سيتم استخدامها بين قوسين):

تحديد موقع المستودع: (1) المكونات الصخرية ذات جودة عالية (بيانات السجلات + البيانات الأساسية)؛ (2) النفاذية الملائمة؛ (بيانات السجلات + البيانات الأساسية)؛ (3) امتداد المستودع (بيانات السجلات + البيانات السيزمية)؛ (4) الصافي إلى النسبة الإجمالية (بيانات السجلات + البيانات الأساسية).
تحديد موقع مانع التسرب: (1) موانع تسرب المكونات الصخرية نموذجية النوع (بيانات السجلات + البيانات الأساسية)؛ (2) الطبقات النفاذية المنخفضة (بيانات السجلات + البيانات الأساسية)؛ (3) امتداد موانع التسرب فوق المستودع (بيانات السجلات + البيانات السيزمية)؛ (4) السُمك الكافي (بيانات السجلات + البيانات السيزمية).
تحديد المصايد: (1) وجود الهياكل (البيانات السيزمية)؛ (2) وجود مصائد استراتجرافية (البيانات السيزمية)؛ (3) وجود إغلاقات رباعية الانخفاض (البيانات السيزمية)؛ (4) وجود إغلاقات رباعية الانخفاض (البيانات السيزمية).

لذلك، يتمثل الهدف من البحث في استخدام تحليلات البيانات الهائلة المتاحة وتقنيات وأدوات الذكاء الاصطناعي المتاحة من صناعة النفط والغاز في المنبع للمساعدة في تحديد مواقع المستودعات المحتملة لعزل ثاني أكسيد الكربون المحتملة عن طريق استخدام بيانات مثل بيانات السجلات والبيانات الأساسية والبيانات السيزمية.

التنبؤ بنفاذية التكوين التي يتم الحصول عليها من السدادات الأساسية في بئر استكشافي محفور في حقل "فولف" الذي يقع في الجرف القاري النرويجي عن طريق استخدام نموذج شبكة عصبية اصطناعية أحادية الطبقة. وكانت بيانات الإدخال التي تمت تغذيتها في الشبكة عبارة ثلاثة سجلات شائعة تتضمن سجل أشعة جاما ((GR، وسجل الكثافة الكبيرة (RHOB)، وسجل المسامية النيوترونية (NPHI). وقد تم تحديد نموذج الشبكة العصبية الاصطناعية أحادية الطبقة الأمثل بأنه يحتوي

على 18 خلية عصبية في الطبقة الخفية، ووظيفة نقل أحادية الخلايا. ووجد أن العدد الأمثل لنقاط البيانات المختارة عشوائيًا من مجموع 557 نقطة بيانات هو 225 نقطة.

ثم استخدمت البيانات التي تم الحصول عليها من مواصفات تكوين الصخور في سجلات نفس البئر في دراسة منفصلة تهدف إلى تصنيف تكوين الصخور. ويعد استخدام الشبكة العصبية الاصطناعية أحادية الطبقة في تصنيف تكوين الصخور مهمًا لتحديد مواقع عزل ثاني أكسيد الكربون المناسبة لأن علم الدهون يلعب دورًا رئيسيًا في خصائص الصخور (مثل النفاذية والمسامية) التي ستكون موجودة. وهذه الخصائص مهمة لإتاحة تحرك ثاني أكسيد الكربون إلى حيز المسام بمجرد حقنه في التكوين عن طريق آبار الحقن.

تم التنبؤ أيضًا بالقيم المسامية أيضًا من بيانات سجل البئر من خلال استعمال نموذج شبكة عصبية اصطناعية أحادية الطبقة. تم استخدام ثلاثة سجلات إدخال وكانت على النحو التالي: أشعة جاما ((GR، والكثافة الكبيرة ((RHOB، والمسامية النيوترونية. وتم ربط قيم السجل بقيم المسامية الأساسية المقابلة التي تم الحصول عليها من التحليل الأساسي لعدد 664 سدادات أساسية. تنقسم مجموعة البيانات إلى مجموعتين: تدريب بنسبة 70% واختبار بنسبة 30%. ويتحقق تحسين النتائج عن طريق مقارنة ثلاثة بارامترات إحصائية رئيسية تشمل ما يلي: معامل الارتباط ((CC، وخطأ الجذر التربيعي (RMSE) ومتوسط النسبة المئوية للخطأ المطلق (AAPE). وقد وجد أن العدد الأمثل من النيوترونات هو 18، حيث تم استخدام وظيفة تحويل الدالة الأسية في الشبكة. وكان معامل الارتباط 0.808، وكان خطأ الجذر التربيعي 0.055، وكان متوسط النسبة المئوية للخطأ المطلق 5.94. وتوضح النتائج أنه يمكن استخدام الشبكة العصبية الاصطناعية أحادية الطبقة كأداة للتنبؤ بالمسامية في المناطق غير المستكشفة من أجل وصف موقع عزل ثاني أكسيد الكربون. وتلعب المسامية عاملاً رئيسيًا في تحديد كمية ثاني أكسيد الكربون التي يمكن تخزينها في موقع معين.

1. CHAPTER 1 – INRODUCTION AND RESEARCH SIGNIFICANCE

1.1. Research Context

1.1.1. Big Data

Big Data is a widely used term that shows up in a variety of contexts in the past few years. The definition of the term, however, may be quite unclear to several people. The meaning that first pops into mind is a large quantity of data that needs to be analyzed or a process that would generate large amounts of data. While this definition is somewhat true as regards to the *volume* aspect of the data, the term, however, encompasses a much larger definition. Another question that comes to mind is how big is big? The term may also be subjective in its nature. What one organization or institution regards as big might not be the same with other larger and better equipped organizations. Therefore, before moving on to the details of this research, let us first define big data.

Big Data is often defined in terms of a number of V's. The definition first started off with 3 V's which were coined by Doug Laney of Gartner (Workflows for Data Science Center of Excellence, 2018). These V's were *volume*, *velocity*, and *variety*. Subsequently, other authors started adding more V's to the definition as this field of study became more popular amongst the scientific and business communities. Other more recently added V's include *veracity*, *variability*, and *value*. These V's were effectively used to describe several features that Big Data incorporates (Workflows for Data Science Center of Excellence, 2018).

- The *volume* of the data refers to the quantity of the data that is generated per unit time. The unit of time is typically in units of seconds, minutes, hours, or days. Depending on the type of operation or process, the amount of data that is generated could indeed be very large. For instance, according to statistics compiled in 2013, every minute we send 204 million e-mails, generate 1.8 million Facebook likes, send 278 thousand Tweets, and upload 200,000 photos to Facebook (Qmee, 2013).
- The *velocity* of the data is used to describe the speed or rate in which the data is generated. It is also used to refer to the pace at which the data travels from one place to another. For instance, Google alone processes on average over 40 thousand search queries per second, making it over 3.5 million in a single day (InternetLiveStats, 2018).
- The *variety* of the data refers to the several different types and forms of data that can be stored digitally. Examples include text messages such as tweets and comments on an online video, video content on YouTube, audio files such as music and podcasts, GPS data, etc.
- The *veracity* of the data refers to the varying quality in which you can receive data. There are several reasons as to why the quality may differ. For instance, the quality of temperature data could be very bad and unreliable due to a faulty thermostat that acts as the sensor. The data in itself may also have a lot of uncertainty attached to it. For instance, the personal information entered by users of an online gaming platform.

- The *variability* of the data can generally refer to the different types of data that one can obtain. This may be in the form of data coming in different dimensions, the differences in terms of whether the data is structured or unstructured, or difference in the velocities in which this data is loaded onto your data storage facility (Firican, 2018).
- The *value* of the data refers to the business benefit that analyzing and processing complex and large datasets can bring about to an organization.

With such massive amounts of data generated on a daily basis, there is indeed great interest in learning how to properly manage, model, analyze, and process these datasets to bring in useful insights that may be used to better understand the world we live in and help us make more informed business decisions. These complex datasets would sometimes require a lot of CPU resources in order to properly analyze them. As a result, new innovative technologies have emerged to tackle these problems, and open-source projects such as Apache™ Hadoop® have been launched.

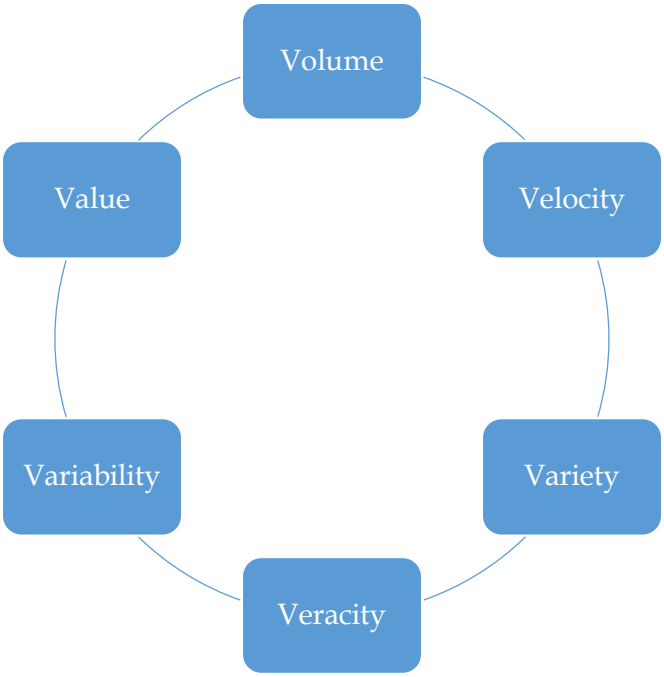


Figure 1.1 Summary of the 6 V's used to describe Big Data.

A great deal of the Big Data technologies available today are based on the Apache™ Hadoop® project (Baaziz & Quoniam, 2014) that was launched by Yahoo! in 2005 following the release of Google's in-house processing framework called MapReduce. Following the release of Hadoop, other frameworks and tools were released to the general public as open-source projects. These frameworks provided new capabilities missing in Hadoop, such as SQL querying or high level scripting. Over the years several open-source projects were made available, many of which rely on Hadoop. In order to understand the relationships between all these frameworks and what each is capable of doing, we may

make use of a layer diagram. In a layer diagram, a component uses the functionality or capabilities of the components in the layer below it. Usually components at the same layer do not communicate, and a component never assumes a specific tool or component is above it. ***Figure 1.2*** represents one set of tools in the Hadoop ecosystem as a layer diagram (Altintas, 2018).

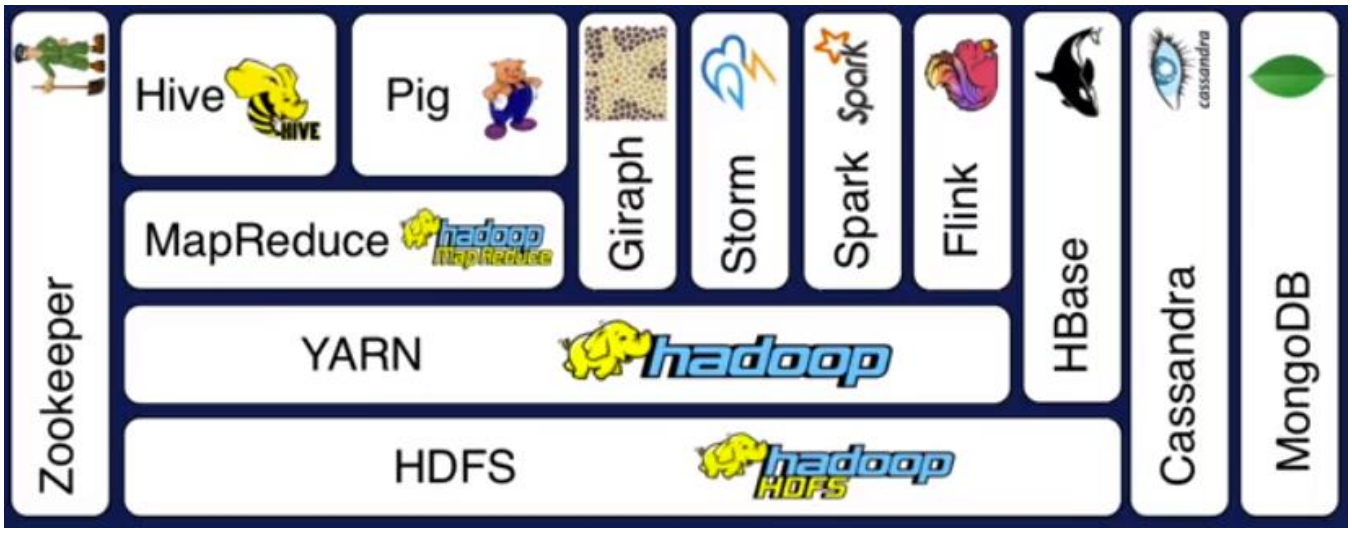


Figure 1.2 A layered diagram representation for Hadoop *(Altintas, 2018).*

This layer diagram is organized vertically based on the interface. Low-level interfaces such as storage and scheduling are located on the bottom, and high-level languages and interactivity are located on the top (Altintas, 2018).

Now that we have defined *Big Data*, let us look into the possible applications Big Data analytics may offer in the context of locating suitable CO_2 sequestration reservoirs. But first, a look into Big Data applications in a number of industries, and a brief introduction to the concept of storing CO_2 underground in geological reservoirs (CO_2 sequestration).

1.1.2. Big Data Applications in Other Industries

Big Data has had a range of applications in different business scenarios. Many industries have sought to use Big Data analytics to gain further insight into their data, make better and more informed decisions, and ultimately transform their businesses. Big data applications have been used to improve security, understand consumer behavior, boost sales, improve healthcare services and well-being, reduce fraud, and improve municipal infrastructures (Marr, 2015).

In the hotel retail business, Big Data is being used to better evaluate the feedback given by customers via analysis of comments and reviews posted on social media sites such as Twitter, Facebook, etc. The technique is known as sentiment analysis and it can provide the hotel management with information regarding the percentage of positive, neutral and negative comments customers post online. This form of analytics will revolutionize the way hotels can deliver and improve the services they provide. The technique is seen as a

more reliable, cost-effective, and efficient way to obtain feedback from customers as opposed to the traditional method of handing out surveys to customers (Marr, 2015).

In retail shops and grocery stores, Big Data analytics can be used to better understand consumer behavior. For instance, many shops and stores are equipped with CCTV cameras for the primary purpose of security. Big Data analytics, however, offers these businesses the chance to observe how their customers behave while they roam around the shop via the use of the same CCTV cameras. Data such as this can provide the business with information such as where to customers usually stop to look at products, which areas are rarely visited, how long to they spend looking at a particular product, etc. Knowing this, the business may decide to revise their showcase design, reduce waiting time, and direct customers to where they need to go to find what they are looking for (Marr, 2015).

Finally, Big Data analytics can also be used in the sports industry to improve performance and physical well-being. An Olympic cycling team has made use of sensors to monitor the performance of cyclists during a race or training session. The sensors were fitted to the pedals of the bicycles and would measure how much acceleration and thrust each cyclist would generate. The team also integrated performance data with health data such as calorie intake, sleeping patterns, and heart rate. With all this information the team was able to successfully make incremental improvements to the team's performance in the Olympics (Marr, 2015).

1.1.3. CO_2 Sequestration

It is now recognised by a number of scientists that global temperatures have been on the rise in the past decades due to rapid industrialization of the modern economy (Baines & Worden, 2004). The industrial revolution has brought forth with it the extensive consumption of sources of energy that rely on the burning of fossil fuels such as coal, natural gas and oil. The exponential population increase in the past few centuries means that more people will be in need of consuming these sources of energy in order to meet their basic needs and uphold their ever increasing standards of life. These factors not only put excessive strain on our natural non-renewable energy resources, but also leads to an increase in the pollution on Earth. The major form of pollution that is caused by the burning of fossil fuels is the release of the well-known green-house gas, carbon dioxide (CO_2). CO_2 is the main compound that has been identified to be affecting the stability of Earth's climate. It represents 62.5% of all green-house gases generated globally (Baines & Worden, 2004). Due to the heating effects that CO_2 has been shown to pose to the Earth, several researchers and institutions have attempted to curb CO_2 emissions via the use of ingenious methods of mitigation. One of these methods, that is thought to have great potential, includes the capture and storage of CO_2, also known as CO_2 sequestration, in deep geological formations such as depleted oil and gas reservoirs, unmineable underground coal seams and deep non-potable water aquifers (Baines & Worden, 2004). These geological storage options are summarized and illustrated in ***Figure 1.3***. In order to inject the CO_2 into underground reservoirs, one first needs to capture the CO_2 from the sources that emit it. There are two main approaches to capturing CO_2 from the atmosphere: (1) capturing it directly from industrial sources. This may include (i) post-combustion capture, (ii) pre-combustion capture, or (iii) burning fossil fuels in a pure oxygen

environment so as to only produce CO_2 as a by-product; (2) capturing CO_2 from the atmosphere by use of natural biological processes which work to fix carbon in plants, marine sediments, and soils (Benson & Franklin M. Orr, 2008).

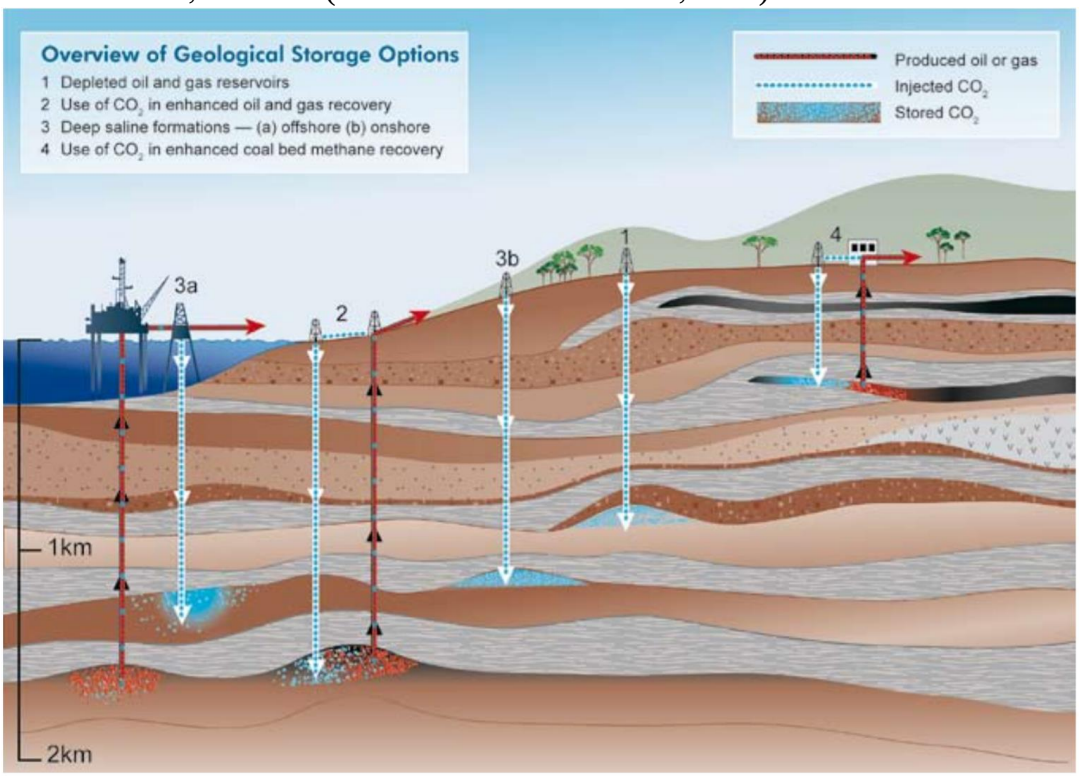


Figure 1.3 Overview of geological storage options for CO_2 (Intergovernmental Panel on Climate Change, 2005).

The injection of CO_2 into oil and gas reservoirs is already being practiced for the purposes of enhancing oil and gas recovery (CO_2-EOR and CO_2-EGR). In the case of oil recovery, the supercritical CO_2 helps to dissolve the oil that is generally left behind in the reservoir (residual oil) and hence mobilize it for extraction. In the case of gas recovery, the injected CO_2 helps to maintain the reservoir pressure above the dew-point pressure of the natural gas phase, hence preventing valuable intermediate hydrocarbon components from condensing in the pore space. The condensation of intermediate components, or condensates as they are more commonly known, is an undesirable process since the recovery of these components would be hindered. Proponents of the geological storage of CO_2 have pointed out that the oil and gas industry offers a representative analogue to the storage of CO_2 underground. The same technology may be utilized for the purpose, however the disadvantage in this case would be that there would be little economic value in solely storing CO_2 without enhancing the production of hydrocarbons, which is a valuable commodity. To this end, a number of authors have proposed the idea of injecting CO_2 to simultaneously enhance hydrocarbon recovery and store it underground for the sake of mitigating global warming and climate change. The estimated additional costs for generating electricity from a power source that utilizes CCS technology to reduce its carbon footprint range from around $20 to $70 per tonne of CO_2 removed from process effluent. The cost would vary depending on the type of capture and storage technology that is being utilized. It is hope that, in the future, R&D efforts would further help to reduce the costs of CCS endeavours, hence, improving the viability of the technology to be applied to a global

scale (Benson & Franklin M. Orr, 2008). Moreover, with the onset of new carbon tax regulations imposed by international governmental organisations and local governments, the storage of CO_2 may indeed work in favour of a nation's economy. The global storage capacity for oil and gas reservoirs has been estimated and is compared with the projected total emissions between 2000 and 2050 according to IPCC's ''business as usual'' scenario. According to the estimates, these reservoirs are capable of accepting 45% of the CO_2 that needs to be stored, a figure that corresponds to 920 Gt of CO_2 (Gale, 2004).

In areas having a wealthy distribution of underground unmineable coal seams, the injection of CO_2 into these coal reservoirs may serve as a viable option. Authors have pointed out that the storage capacity of coal seam reservoirs is relatively low in comparison to the other two reservoir types. According to IPCC's estimates, unmineable coal seams may accept less than 2% of the CO_2 that needs to be stored, a figure corresponding to only 20 Gt of CO_2 (Gale, 2004). The mechanisms whereby the CO_2 is trapped in coal seam reservoirs include two main types: the adsorption of CO_2 on the coal surface that would thereby displace adsorbed methane gas, CH_4, and the physical trapping in the cleats within the coal itself (Gale, 2004). In this way, the CO_2 would be stored in much the same way as hydrocarbon gas is stored for several million geologic years within the coal seams. In order for this solution to prove viable, it is important that the coal mine be abandoned and not mined again.

Finally, there are several research groups that sought to study the storage of CO_2 in saline water aquifers or aquifers that are unusable for the purposes of drinking and domestic purposes, for instance. One issue of concern that was brought up regarding the storage of CO_2 in water aquifers was the reaction of CO_2 with the water to produce the weak carbonic acid which may in turn react with carbonate rocks in the subsurface. This issue has been tackled by a number of authors, and there seems to be a consensus that the injected CO_2 does not react significantly with minerals present in the formations of sandstone and carbonate reservoirs. The principle whereby the CO_2 is stored and trapped in the reservoir is as follows:

- Due to buoyancy effects the injected CO_2 would seek to rise to the surface. This is analogous to natural gas migrating upwards after it is generated from a gas prone source rock and accumulating on top of the oil in the pore space.
- The rising CO_2 would then be hindered from further migration by the presence of a suitable sealing layer. Common sealing layers include shale, silt and evaporite type rocks such as halite, gypsum and anhydrite. The permeability and thickness of a sealing layer need to be sufficient for the sufficient accumulation of the injected CO_2.
- Over time, the CO_2 will then dissolve into the formation water—the rate of dissolution will be controlled by the surface area of CO_2 in contact with the formation water (Gale, 2004).

Authors have sought to understand the mechanisms that take place upon injection of CO_2 in saline aquifers in order to gauge the long-term fate of the injected gas. A scaling analysis of the convective mixing of CO_2 in saline aquifers was studied via the use of a numerical simulation (Hassanzadeh, Pooladi-Darvish, & Keith, 2007). It should be noted, however, that the storage capacity for saline water reservoirs is uncertain to some extent. The

estimates range from 20% to 500% of the CO_2 that needs to be stored. These figures correspond to 400 to 10,000 Gt of CO_2 (Gale, 2004).

Pilot tests are already underway to evaluate the viability and feasibility of CO_2 storage in the above mentioned geological reservoirs. In a gas field located in the North Sea, CO_2 has been injected into a depleted gas reservoir in a sand sequence[1] known as the Utsira formation. The injection process began in 1996, and nearly 8 million tonnes of CO_2 have been injected into the formation by the writing of the paper that was following this major project. The results seem promising, as no significant operational problems were observed in both the process of capturing the CO_2 and injecting it via the injection wells (Torp & Gale, 2004). The authors, however, point out that it must be demonstrated that CO_2 storage is both safe and has a low environmental impact.

The idea of injecting waste products into deep underground reservoirs is not a new one. Industrial and radioactive wastes have been disposed of in deep reservoirs for decades. The safety of the proposed storage method raises concerns amongst environmentalists and the general public. For instance, people raise concerns over whether the waste product would leak into potable water aquifers, to the surface, or under residential areas (Gale, 2004; West, Pearce, Bentham, & Maul, 2005). Authors have also embarked upon trying to evaluate the possible environmental ramifications of a possible CO_2 leakage from a carbon capture and storage (CCS) project. The issue involves not only the leakage of the CO_2 gas in itself, but also the possible transportation of other hazardous impurities commonly found in the subsurface that include, hydrogen sulfide (H_2S), nitrogen oxides (NO_x) and sulfur oxides (SO_x). Due to the following concerns regarding the possible leakages of CO_2 from CCS projects, the development of the project on a large scale may be hindered by concerned parties. Possible health hazards that may occur in organisms exposed to elevated levels of CO_2 include headache, increased blood pressure, difficulty in breathing, dizziness, sweating, near or full unconsciousness. Excessively high CO_2 exposure concentrations may also lead to a coma and eventual death of humans (West, Pearce, Bentham, & Maul, 2005).

1.1.4. Literature Review: Reservoir Properties for Ideal CO_2 Sequestration Site

Fang *et al*. (2010) provided an overall review of the process of sequestering CO_2 in geological saline water aquifers. The review pointed out several characteristics that are important in classifying a suitable candidate for sequestration. It is well documented in the literature that saline water aquifers possess a great capacity for the storage of CO_2. Some of the important characteristics pointed out include: (i) adequate connected porosity, (ii) CO_2 density large enough to ensure economical storage, and (iii) formation injectivity large enough to avoid a large pressure increase when injecting CO_2. Important properties of the basin include: (i) Basin type and tectonic setting, (ii) hydrodynamic and geothermal

[1] It should be noted that the sequestration of CO_2 into carbonate sequences, as opposed to sand sequences, may lead to slight complications in terms of the geochemical reactions that may take place within the reservoir/aquifer. It has been shown experimentally that the injected supercritical CO_2 would alter the pH of the brine to a more acidic state. This lowering in the pH may speed up dissolution processes within the reservoir (Kaszuba, Janecky, & Snow, 2003).

regimes[2], and (iii) basin resources and maturity. Important reservoir properties include: (i) geologic structure, (ii) cap rock integrity, (iii) size, (iv) depth, (v) porosity and permeability, and (vi) fluid properties and rock mineralogy. Chadwick *et al*. (2004) set out to draw some general conclusions about the storage of CO_2 in the Utsira Sand as part of the Sleipner Project present in the northern North Sea. Important properties for a CO_2 sequestration site include: (i) permeability barriers/Intra-reservoir trapping: The shale layers constitute important permeability barriers within the reservoir sand, and have proved to have a significant effect on CO_2 migration through, and entrapment within, the reservoir, (ii) pore volume enclosed within structural and stratigraphic traps, wherein CO_2 can be expected to accumulate in the long-term, (iii) reservoir structure and stratigraphy, and (iv) other small reservoir heterogeneities that may not be apparent in the seismic profiles prior to the injection of CO_2. Kovscek (2002) laid down a general groundwork/screening procedure that can be used to help classify a site and determine whether or not it may serve as a suitable site for CO_2 storage. The author, however, points out that not all the criteria outlined in the procedure need to be fulfilled in order to have a suitable site. Some projects have proved successful even though they did not meet all criterion. Important properties for an ideal site may be divided into three main aspects and include:

- **Reservoir Engineering Aspects**: (i) Carbon density, (ii) specific capacity, (iii) injectivity, (iv) reservoir flow mechanics, (v) aquifer-reservoir coupling, and (vi) incremental oil recovery.
- **Geophysical Aspects**: (i) Seals, faults, and fractures, (ii) formation damage, and (iii) monitoring.
- **Surface Facilities Aspects**: (i) Cost of concentrating CO_2 in dilute waste gas streams, (ii) cost to build infrastructure to transport CO_2, and (iii) corrosion that may be caused by impurities in the CO_2 gas stream.

Bachu (2000) proposed a step-wise approach for assessing a prospective site and selecting the methods for storing the CO_2. The author points out that several criteria have to be considered when evaluating the potential of a sedimentary basin for CO_2 sequestration: (i) its tectonic setting and geology, (ii) the basin geothermal gradient, (iii) the hydrodynamic regime of formation waters, (iv) the hydrocarbon potential and basin maturity, and (v) economic aspects relating to access and infrastructure and socio-political conditions. At a later stage, a continuation of the work by the same author (Bachu, 2002) concluded the following points:

[2] The temperature of the basin has an effect on the amount of CO_2 that can be stored into a reservoir. According Fang *et al*. (2010), colder basins are more favorable to the storage of CO_2 since they permit the storage of dense CO_2 at depths that are relatively shallow. Keep in mind that the density of a fluid increases as the temperature decreases.

i. A series of geoscience, engineering, economic and public issues need addressing by governments and industry before proceeding with full scale implementation of CO_2 sequestration projects.
ii. Suitability analysis of a site is based on geological, geothermal, hydrodynamic, basin maturity, economic and societal criteria.
iii. Physical state of injected CO_2, density, and viscosity affect CO_2 fate, trapping mechanism, and capacity of geological sinks.
iv. Transforming geological space into CO_2 space is an important step.

1.1.5. Literature Review: Locating Reservoirs - A Hydrocarbon Prospect Analysis Approach

The approach we wish to take to locate ideal CO_2 sequestration sites would be akin to locating hydrocarbon prospects. Several data sets need to be analyzed in order to pinpoint the location of a possible hydrocarbon prospect. Several researchers and industry professionals have attempted to summarize the steps involved.

Haris *et al*. (2017) made use of well log, check shot, 3D post-stack seismic and dipmeter data to interpret seismic attributes and inversions. The properties that were important in classifying the reservoir included porosity, sedimentation trend, reservoir distribution and geometry, and the distribution of reservoir porosity. Roisenberg *et al*. (2009) attempted to provide a method of systematizing the process of prospect appraisal. The utilized seismic data analogies and geological theories to determine properties such as NTG ratio values, burial depth, and porosity as regards to reservoir rock properties; thickness, and seal continuity with regards to seal rock; and finally, quality of seismic grid, and type of trap as regards to the trap. The authors conclude that the fuzzy approach is capable of handling incomplete data and imprecise information that commonly characterizes the exploration of hydrocarbons.

Otis and Schneidermann (1997) provided a summary of the exploration evaluation process that has been used by Chevron Overseas Petroleum, Inc. to provide estimates of exploration prospect values. The geologic risk of the prospect is evaluated by taking into consideration the probability that four key factors of the play concept:

i. Existence of mature source rock.
ii. Presence of reservoir rock, which can be determined by looking at properties such as lithology, reservoir distribution, depositional model, lateral continuity and extension, and thickness and vertical cyclicity.
iii. Presence of trap, which can be determined by looking at properties such as lithology and ductility, thickness, continuity, curvature, and degree of fracturing or faulting.
iv. Presence of play dynamics.

The probability of geologic success is then obtained by finding the product of the probability of occurrence of each of the four factors.

Raef *et al*. (2016) utilized several different seismic attributes that included acoustic impedance, amplitude, spectral decomposition, two-way travel time (TWTT) and

amplitude, for the purpose of properly delineating potential hydrocarbon reservoirs. The authors recommended that future exploration endeavors take into consideration the geological channel feature that has been outlined in their study, which may serve as good reservoir quality rock. However, they pointed out that some of the most significant concerns for the prospects identified within the channel would be the possibility for lack of structural development/closure and the possibility of reservoir compartmentalization. Hui *et al*. (2009) pointed out that care needs to be taken when dealing with volcanic rock oil/gas exploration targets, which may be characterized by seismic data with low magnetic data. The authors made use of petrophysical, magnetic, gravity, electric, seismic and drilling data, combined with geological regularities to draw up their conclusions. Oyeyemi *et al*. (2018) made use of suites of digital wireline logs (e.g. gamma-ray, resistivity and density logs), checkshot, and seismic data to characterize and evaluate hydrocarbon resources using integrated well logs analysis and 3D seismic-based reservoir characterization in an offshore field, western Niger Delta basin. Some of the conclusions that were drawn up from their study included: (i) seismically derived chimneys were successful in highlighting hydrocarbon migration pathways, and (ii) application of seismic colored inversion on the seismic volume is capable of revealing reflectors and reservoir units in a sharper and clearer manner. The authors, however, pointed out that it is necessary to provide core data and biostratigraphic data to further their research. Harrison and Kennedy (2002) stressed the importance of incorporating petrophysical model early in the workflow of evaluating prospects. The petrophysical properties that were utilized include porosity, permeability, saturation, relative permeability, and NTG. The authors concluded that neglecting input from a petrophysicist when estimating the potential reserves of an exploration opportunity will lead to major errors in the final prospect evaluation. Spofforth and Firth (2016) highlighted that integrated new interpretation with geology-based datasets will prove key to fully understanding petroleum prospectivity. They concluded that the integration of these datasets offers several advantages, even before the seismic data is acquired. Incorporating knowledge of the geology into the survey design means that the optimum data can be acquired with the ideal offsets and azimuths to image the target.

1.1.6. Big Data in the Oil & Gas Industry: A Representative Analogue

Although the idea of incorporating Big Data analytics into the improvement of the Oil & Gas industry is not a new one, the majority of initiatives are still in the experimental stage (Baaziz & Quoniam, 2014). Oil & Gas service companies, however, employ a great deal of sensors that measure data continuously in real-time and relay the information to decision makers back to base. The type of data that is generated in the upstream petroleum industry, like most industries, may be classified into three main types (Marr, 2015):

- *Structured data*: Data that is located in a fixed field within a defined record or file. These type of data is handled with certain applications that are used to manage surveying, processing and imaging, exploration planning, reservoir modeling, etc. (Baaziz & Quoniam, 2014)

- *Semi-structured data*: These datasets would fall in between structured and unstructured data. Some of the data may be classified into particular known fields, whereas the rest would be unstructured.
- *Unstructured data*: Datasets that are not classified at all. These datasets are the most cumbersome to work with, however, if analysis is done properly, the insights they may bring forth can be highly useful for any business.

In order to make good use of both semi-structured and unstructured data, there is a pressing need to make use of Big Data technologies, especially if the data can be classified by the 6 V's of Big Data.

Examples of companies that are attempting to make use of Big Data tools in the upstream petroleum industry include the following (Baaziz & Quoniam, 2014):

- Chevron proof-of-concept using Hadoop (IBM BigInsights) for seismic data processing;
- Shell piloting Hadoop in Amazon Virtual Private Cloud (Amazon VPC) for seismic sensor data;
- Cloudera Seismic Hadoop project combing Seismic Unix with Apache Hadoop;
- PointCross Seismic Data Server and Drilling Data Server using Hadoop and NoSQL;
- University of Stavanger data acquisition performance study using Hadoop.

In summary, ***Table 1.1*** summarizes the types of Big Data that can be obtained from the Oil & Gas industry and classifies them according the 6 V's that were previously discussed (Baaziz & Quoniam, 2014).

Table 1.1 Big Data in the Oil & Gas Industry.

	Exploration	*Reservoir Engineering & Development*	*Drilling and Completion*	*Production*
Volume	Seismic acquisition SEGD	Facilities Reservoir Engineering	Sensors: Flow Pressure ROP	SCADA Sensors: Flow Pressure
Variety	Structured data: • SEGD • Pre-stack • Post-stack Semi-structured data: • Implantation Report	Structured data: • WITSML (XML) • PRODML • RESML Unstructured data: • Log Curves/Drilling & Test/Lithology/Cores…	Structured data: • WITSML Semi-structured data: • Final Well Report • Daily Drilling Report Unstructured data: • Drilling Log/Gas Log…	Structured data: • PRODML • RESML Semi-structured data: • Crude Analysis Report
Velocity	Real Time Data Acquisition: Wide azimuth data acquisition		Real Time Data Acquisition: Mud Logging/LWD/MWD	Real Time Data Acquisition: SCADA Sensors
Veracity	Seismic processing	Reservoir Modeling	Sensor calibration	Sensor calibration

Variability	Seismic Interpretation Geology Interpretation	Reservoir Simulation Combination of seismic, drilling and production data	Data Interpretation and Optimization	Data Interpretation
Value	Navigation, Visualization and Discovery Run integrated asset models	Improve drilling program Drive innovation with unconventional resources (shale gas, tight oil)	Reduce costs Reduce Non-Productive Time (NPT) Reduce risks Improve HSE performances	Increase speed to first oil Enhancing production

1.1.7. Artificial Intelligence Techniques in the Oil & Gas Industry: A Representative Analogue

Now that we have a better understanding of what *Big Data* entails, we look at how we can use *Artificial Intelligence* algorithms and techniques to derive insights from our immense data set. The two words – *Artificial Intelligence* and *Big Data* – appear very often in the wake of the 4th Industrial Revolution. They can be found anywhere from your local and international news broadcast, in technology magazines, in ground breaking research work across the globe, and in several conference proceedings. The terms are often used interchangeably, however, in this section we shall briefly attempt to shed light on the similarities as well as the differences between the two.

Artificial Intelligence and Big Data are both tools that may be utilized to gain better insights into your data. One way of looking at how these are inherently different from each other is by looking at Big Data as the raw unprocessed data that needs to undergo the cleaning, processing, structuring, sorting, etc. in order for it to be ready for analysis. Whereas, on the other hand, Artificial Intelligence is the algorithms and techniques that are then applied to this processed data set to generate an output that is beneficial. Coupling both of these tools together then helps us gain insights into the data and make more informed decisions as a result (Patrizio, 2018).

The subsequent sub-sections will discuss some of the Artificial Intelligence techniques that are available. The discussion will mainly be limited to Artificial Neural Networks (ANN) and Fuzzy Logic. The final sub-section will look at applications of these tools in reservoir property estimation and prospect evaluation.

1.1.7.1. Artificial Neural Network (ANN)

The fundamental principle of the ANN revolves around the following terms (The MathWorks, Inc., 2009):

- Scalar or vector input
- Weight
- Bias
- Transfer function
- Output

Input is the data that you input into the network. Keep in mind that the quality control (QC) and statistical analysis should have already been performed on this data before uploading

it onto the network. The QC of the data will help us determine any anomalies in the data or unexpected trends. The input is then multiplied by a *weight* which is a scalar quantity. Additionally, the product of the input and the weight is added to a *bias*, which can be viewed as a shift of the *transfer function* that performs the final computation to produce the output. Refer to ***Figure 1.4*** for an illustration of how these terms interact with each other.

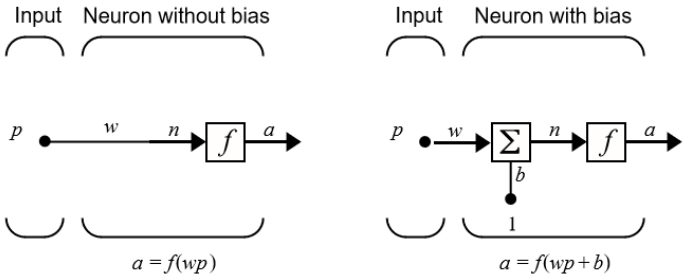


Figure 1.4 The interaction of the different terms used in describing ANN (The MathWorks, Inc., 2009). Where p is the input, w is the weight, wp is the product of the weight and input, b is the bias, and a is the output.

The transfer function is what is used to convert the modified input to the final output. There are several different transfer functions that are available for use. The transfer function that is best to use would largely depend on the type of problem you are dealing with. Examples include:

- Pure Linear
- Tan-sigmoid
- Log-sigmoid

A simple approach would be using a variety of these to determine which will produce the best results.

1.1.7.2. Fuzzy Logic

According to MATLAB®'s Fuzzy Logic Toolbox™ 2 User's Guide (The MathWorks, Inc., 2010):

> *"Fuzzy logic has two different meanings. In a narrow sense, fuzzy logic is a logical system, which is an extension of multivalued logic. However, in a wider sense fuzzy logic (FL) is almost synonymous with the theory of fuzzy sets, a theory which relates to classes of objects with unsharp boundaries in which membership is a matter of degree. In this perspective, fuzzy logic in its narrow sense is a branch of FL. Even in its more narrow definition, fuzzy logic differs both in concept and substance from traditional multivalued logical systems."*

1.1.7.3. Literature Review: Applications in Reservoir Property Estimation and Prospect Evaluation

Several authors have investigated the use of Artificial Intelligence (AI) techniques in the upstream petroleum industry as well as in the exploration phase. AI techniques such as ANN, BPNN and Fuzzy Logic have been used to try and improve estimations and

predictions of reservoir properties such as porosity and permeability by use of well log and seismic data. Moreover, they've also been used to determine litho-facies and identifying surface phenomena that may be indicative of the presence of oil and gas.

The whole process of hydrocarbon exploration includes several subjective conclusions whose success rate would be largely dependent on the experience of the exploration geologist or geophysicist. AI may be used to reduce the need of things like 'gut feeling' by making use of the vast amounts of data that are generated during the exploration phase of a project. Aminzadeh (1994) points out that there exists a great deal of subjectivity in the exploration of hydrocarbon reserves. The use of fuzzy logic to counter this may help us solve the problem, however this comes with several challenges of its own. The challenges stem from the fact that designing such a complicated AI system would require that all the rules, know how, and information utilized to locating reserves be incorporated to train the AI system. The author illustrates that AI techniques may be used to (i) segment seismic sections based on texture, (ii) integrate data from different disciplines, and (iii) help in stratigraphic interpretation.

Ahn *et al*. (2018) set out to develop a pre-trained neural network that would be able to execute reliable data-analytics that makes use of an ANN structure with a stacked autoencoder (ANN-SAE) to model heterogeneous channel reservoirs. The authors were able to achieve this by generating a total of 200 reservoir models and thereafter extracting the permeability data of each model. Production history data was then extracted and compressed by use of the SAE. The neural network structure was constructed and trained by the use of 160 randomly selected reservoir models, then new models were generated by use of the SAE decoder. The performance of the ANN-SAE system was tested by the use of the remaining 40 reservoir models. Abdulraheem *et al*. (2007) applied fuzzy logic modeling to predict reservoir permeability from conventional open-hole logs. A MATLAB software was coded to make use of fuzzy logic to handle the training of the model and then predict the permeability for either new layers or other wells in the surrounding vicinity. The authors made use of 70% of the data set to train the model, and then tested the model on the remaining 30% of the data set. A study was also conducted to investigate the minimum number of training data points that would allow for a sufficiently accurate permeability prediction. Saggaf (2002) described an approach based on fuzzy logic for estimating reservoir quality by integrating several attributes from multiple input sources. The approach was then applied to compute the reservoir quality in the Haradh area of the Ghawar field located in the Kingdom of Saudi Arabia. Esmaili and Mohaghegh (2016) utilized a data-driven reservoir model in order to model the Marcellus shale asset. The authors aimed to develop an AI-based model that would be able to overcome several current challenges that are associated with shale gas reservoirs. The input data included production history, geomechanical and geochemical properties, and hydraulic fracturing variables. The optimal model was obtained by using 80% of the data for training and the remaining 20% for testing (calibration and verification).

Olatunji *et al*. (2015) attempted to cultivate the unique attributes of a hybrid intelligence system known as the type-2 fuzzy logic system (type-2 FLS) in order to improve the generalisation capabilities of the Sensitivity Based Linear Learning Method (SBLLM). The authors point out that there is inherently a significant amount of uncertainty in the modeling and prediction of reservoir parameters such as permeability and PVT properties. The type-2 FLS is therefore introduced in order to handle and model the uncertainties present in the problem. In the case of the empirical studies, 70% of the data set was utilized to train the model, and then the model was tested on the remaining 30% of the data set. Results point to better performance of the hybrid system in comparison to other commonly used individual models. Taghavi *et al*. (2005) used and compared three different methods to estimate reservoir permeability from well log data. The first method was a relationship between permeability and effective porosity, where calibration was obtained by use of core porosity and permeability values. Second was an extension of the first model by use of multilinear regression. Lastly, the authors made use of fuzzy logic and noted that fuzzy logic offered the best estimation of permeability in comparison to the other two methods. Anifowose *et al*. (2013) also made use of three methods to predict reservoir permeability, however in this case the authors integrate 3-D seismic data as well as log measurements to get a better estimate. The three methods utilized were ANN, Support Vector Machine Learning (SVM), and type-2 fuzzy logic. The data was trained and tested in the 70:30 ratio. The performance of the AI models was quantified and compared by use of statistical measures such as correlation coefficient, root mean-squared error, and mean absolute error, along with the execution time. The authors conclude that combining seismic and log data shows better results in estimating reservoir permeability in comparison to using only one of these datasets for the same purpose.

Aminzadeh and Brouwer (2006) used neural networks in parallel with fuzzy logic in order to improve reservoir property prediction and decrease the risk of a hydrocarbon prospect. The authors conclude that this method helps in high-grading these prospects. Cuddy (1997; 2000) used fuzzy logic to predict the litho-facies (e.g. aeolian, fluvial, sabkha) and permeability of reservoir rocks present in the North Sea. The predicted parameters were compared with core descriptions of the lithology as well as core derived permeability. Results show that the technique can predict litho-facies with up to 73% accuracy in one case. The author concludes that the technique may be used as a simple prediction tool in uncored wells. Malvić *et al*. (2010) presented three case studies from Croatia on the use of NN in the analysis of geological data (obtained via well logs and cores) from hydrocarbon reservoirs. Results of the study point to the capability of the method being used to better understand clastic reservoirs in the study area. Nashawi and Malallah (2010) used FL to predict formation permeability from readily available well log data. Their approach followed a two-step process where the first mined the data, using algorithms, to determine the different rock types present in the formation. The AI systems were then used to predict the permeability (k). The procedure is summarized in the following 4 steps: (1) principal components analysis, (2) cluster analysis, (3) discriminant analysis (4) FL model applied to each cluster to calculate k. A total of 1140 data points were used to train and test the AI

systems, where 375 data points were used for validation purposes. The authors conclude that:

i. the model-based clustering analysis is efficient in classifying the data into different clusters,
ii. discriminant analysis may be used to locate clusters,
iii. FL has significant potential in predicting permeability, which can be compared very well with core permeability data.

Lim and Kim (2004) attempted to determine reservoir porosity and permeability from well log data by use of both FL and NN. The authors conclude that this integrated approach provides more accurate and reliable estimations in comparison with conventional methods.

1.2. Research Objective

The research objective is to identify potential CO_2 sequestration reservoirs by utilizating the available Big Data (i.e. well log data, core data and seismic data) analytics and Artificial Intelligence techniques from the Upstream Oil & Gas Industry.

The research problem can be split into three main sections which can be further subdivided into subsections (data to be used have been put in between parentheses):

Locating the Reservoir:	**Locating the Seal**:	**Locating the Trap**:
•Good Quality Lithologies (log + core); •Favorable Permeability (log + core); •Reservoir Extension (log + seismic); •Net-to-Gross Ratio (log + core).	•Typical Seal Type Lithologies (log + core); •Low Permeability Layers (log + core); •Seal Extension over Reservoir (log + seismic); •Sufficient Thickness (log + seismic).	•Presence of Structures (seismic); •Presence of Stratigraphic Traps (seismic); •Presence of Four-Way Dip Closures (seismic); •Presence of Three-Way Dip Closures (seismic).

However, due to time constraints, the scope of this thesis work will only cover the first section (**Locating the Reservoir**). Any research undertaken beyond this section would be considered as either supplementary or extra work.

1.3. Research Question

The following is the main question that this research would seek to address:

- Can Big Data analytics tools be used to manipulate well log data and seismic data to help in the identification of potential CO_2 sequestration reservoirs?

1.4. Research Methods

1.4.1. Data Set Used

The data set used is provided by *Equinor* from the Volve field located in the Norwegian continental shelf. The data is publicly available for the purpose of facilitating learning and research for students, scientists, and academic professionals. Refer to the extract below (Equinor, 2018) for a summary of the company's initiative and general background information about the Volve field.

> *"As part of our goal of shaping the future of energy, we wish to support the energy innovators of the future. That's why we're sharing a complete set of data from the Norwegian continental shelf, approximately 40,000 files from the Volve field, which was in production from 2008 to 2016. The data have been released to give students and scientists a realistic case to study. Our aim is to support learning, innovation and new solutions for the energy future.*
>
> *Volve is a decommissioned field in the central part of the North Sea and was discovered in 1993, the plan for development and operation (PDO) was approved in 2005. The field was shut down in 2016 and the formal removal and decommissioning will be completed by the end of 2018.*
>
> *The Volve field is located five kilometers north of the Sleipner Øst field with water depths of 80 meters in the block 15/9. The produced oil was from sandstone in the Hugin Formation of the Jurassic age. The reservoir is located at 2750 – 3120 m depth.*
>
> *The Volve Data was approved for data sharing in 2018 by the initiative of the last Operating company, Equinor and approved by the license partners ExxonMobil E&P Norway AS and Bayerngas Norge AS in the end of 2017.*
>
> *During the life of the Volve license, a limited amount of field-related data has been released to a handful of research institutes.*
>
> *The Volve Data available will contain data covering data in regards of production data, well design, completion string design, seismic data, well logs (petrophysical and drilling), geological and stratigraphical data, static and dynamic models, surface and grid data."*

1.4.2. Research Methodology

1.4.2.1. Initial Statistical Analysis

The first step in solving this problem will begin with an initial analysis of the data in order to understand its general behavior. The statistical analysis of data is a process with several phases, each with its own goal. Moreover, statistical analysis provides a summary of data in the form of graphics and parameters, provides critical information to be used in estimation and conditional simulation, and serves as an efficient vehicle for communicating information. The following statistical measures (major parameters summarized in ***Figure 1.5***) and graphic representations shall be used to understand things like the quality of the data, the quality of measurements, and the characteristics of data sample (Montgomery & Runger, 2014):

i. **"Sample Mean**: This is essentially the arithmetic mean of our data set. Since we almost always think of our data as a sample, it shall be referred to as the sample mean.

If the *n* observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample mean is

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (1.1)$$

ii. **Sample Geometric Mean**: The geometric mean of a set of *n* positive data values is the n^{th} root of the product of the data values; that is

$$\overline{x_g} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \qquad (1.2)$$

iii. **Sample Harmonic Mean**: The harmonic mean of a set of data values is the reciprocal of the arithmetic mean of the reciprocals of the data values; that is

$$\frac{1}{\overline{x_h}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i} \qquad (1.3)$$

iv. **Sample Mode**: The mode of a sample is that observed value that occurs most frequently. There may be more than one mode of either a sample or a distribution.

v. **Sample Median**: The median of a set of data is that value that divides the data into two equal halves.

vi. **Sample Quartiles**: The three values (Q_1, Q_2, and Q_3) of a variable that partition it into four equal parts. The central value is usually called the median and the lower and upper values are usually called the lower and upper quartiles, respectively.

vii. **Interquartile Range (IQR)**: The difference between the third and first quartiles (Q_3-Q_1) in a sample of data. The interquartile range is less sensitive to extreme data values than the usual sample range.

viii. **Sample Variance** and **Standard Deviation**: The variability or scatter in the data may be described by the sample variance or the sample standard deviation.

If $x_1, x_2, \ldots, x_n$ is a sample of *n* observations, the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \qquad (1.4)$$

The sample standard deviation, *s*, is the positive square root of the sample variance.

ix. **Range**: In addition to the sample variance and the sample standard deviation, the sample range, or the difference between the largest and smallest observations, is often a useful measure of variability.

If the *n* observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample range is

$$r = max(x_i) - min(x_i) \qquad (1.5)$$

x. **Frequency Distribution**: An arrangement of the frequencies of observations in a sample or population according to the values that the observations take on.

xi. **Histogram**: A visual display of the frequency distribution. The histogram is a univariate data display that uses rectangles proportional in area to class frequencies to visually exhibit features of data such as location, variability, and shape.

xii. **Skewness**: A term for asymmetry usually employed with respect to a histogram of data or a probability distribution.

xiii. **Kurtosis**: A measure of the degree to which a unimodal distribution is peaked.

xiv. **Box Plot**: A graphical display of data in which the box contains the middle 50% of the data (the interquartile range) with the median dividing it, and the whiskers extend to the smallest and largest values (or some defined lower and upper limits).

xv. **Scatter Diagram**: A diagram displaying observations on two variables, *x* and *y*. Each observation is represented by a point showing its *x*-*y* coordinates. The scatter diagram can be very effective in revealing the joint variability of *x* and *y* or the nature of the relationship between them.

xvi. **Probability Plot**: A scatter plot used to judge if data can reasonably be assumed to follow a particular probability distribution. A normal probability plot is often used to evaluate the normality assumption of data or residuals."

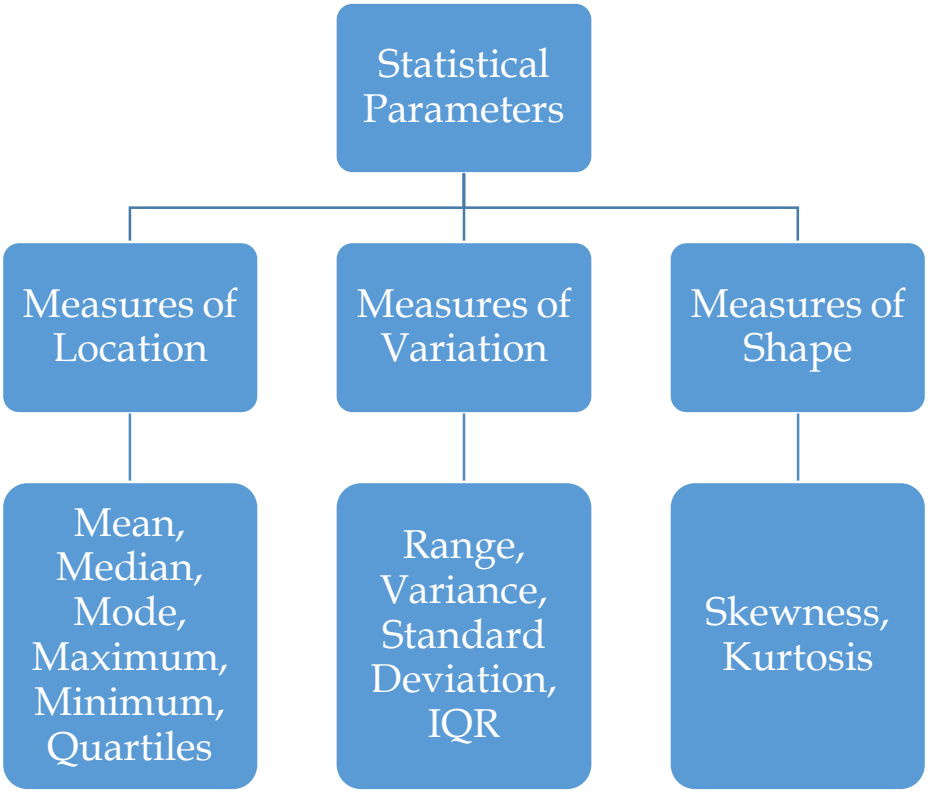


Figure 1.5 Summary of parameters to be measured in the initial statistical analysis.

1.4.2.2. General Guidelines

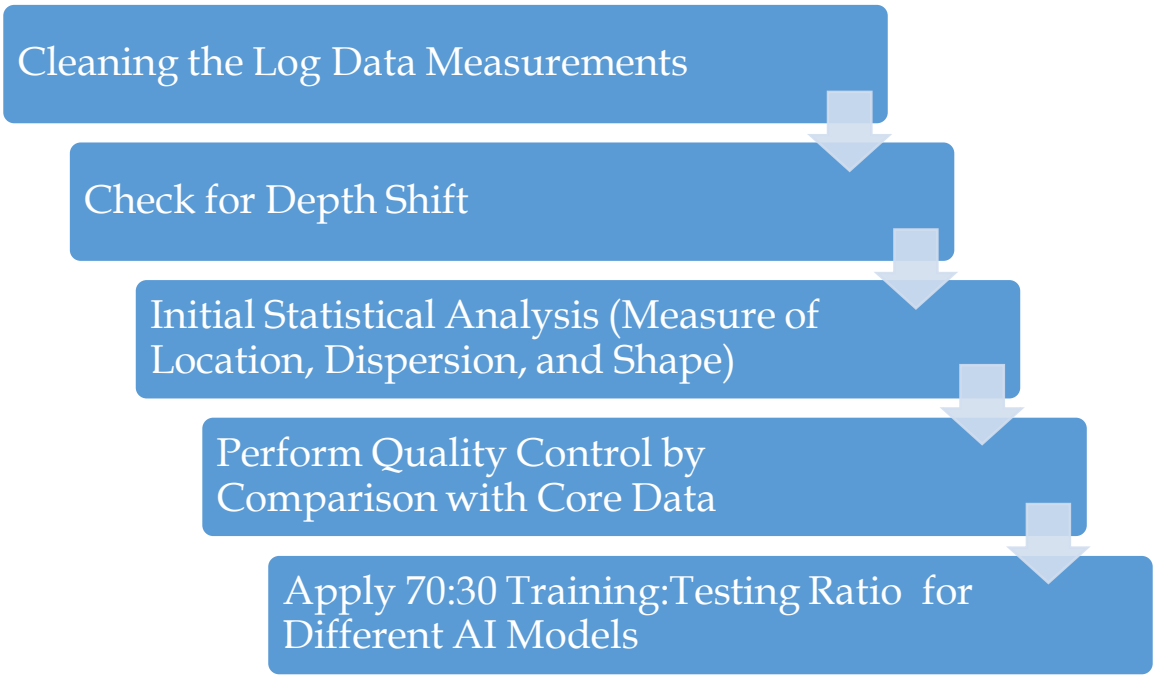


Figure 1.6 Simplified workflow for preparing the log data measurements for analysis by different AI models.

1.5. Significance of Research

1.5.1. Environmental Value

As previously mentioned, the presence of excessive amounts of CO_2 in the atmosphere is thought to be the leading factor that contributes to rising of global temperatures (global warming). The main purpose of carbon sequestration projects is to reduce the total amount of CO_2 that is present in the atmosphere, and hence reduce the heating affect it is believed to cause. A large amount of CO_2 currently present in the atmosphere is thought to be of anthropogenic origin. The atmospheric concentration of CO_2 is thought to be increased by the onset of industrial revolution, and the extensive burning of fossil fuels.

The ideal course of action that is suggested by the public and governments is to shift energy sources from non-renewable to more sustainable renewable sources. However, such a dramatic shift is rather difficult to implement on large scale. Therefore, CO_2 sequestration is proposed to reduce the net CO_2 released into the atmosphere. The process of carbon absorption is one that is present naturally in the environment. Several natural carbon sinks exist, that include plants, soils, and oceans. However, due to the onset of anthropogenic sources of pollution, the balance of CO_2 in the atmosphere has been upset. Now artificial means of storing the carbon are needed to maintain the balance (HelpSaveNature, 2017).

The following are some of the advantages that carbon sequestration may provide: (i) prevention of climate change by reducing the net amount of CO_2 in the atmosphere, (ii) alongside the removal of greenhouse gas from the atmosphere, the injection of CO_2 into geological formations may also be used to enhance the recovery of hydrocarbon resources, (iii) it may be possible to reduce emissions by 80 to 85% by carbon sequestration even

while we are still using fossil fuels, and (iv) as to our knowledge there have been no reports of CO_2 leaking from sites where it has been injected (HelpSaveNature, 2017).

1.5.2. Economic Value

Although the costs of implementing a CCS project may be quite high initially (Hardisty, Sivapalan, & Brooks, 2011), there exist a number of economic benefits as well. According to Yelebe and Samuel (2015), the implementation of CCS technologies in Nigeria has the potential to provide several economic benefits. These benefits include the following: (i) the CCS projects in Nigeria have been estimated to be worth around $10 billion per year by the year 2030, and it is estimated that these projects may generate up to 100,000 additional jobs; (ii) these jobs are especially prevalent for people living within the vicinity of a capture plant, pipeline or storage facility; (iii) CCS is also being proposed as a method of helping in the transition from non-renewable energy sources towards more sustainable sources of energy. The authors state that in the future several energy intensive manufacturing industries will prove to be unsustainable without the intervention of CCS; (iv) in Nigeria, CCS presents a greater potential in helping the country reduce its emissions by at least 25% in accordance with its goal for 2020.

A study conducted by the U.S. Department of Energy (2016) outlined the use of CCUS technology in the U.S. It points out that these technologies provide a key solution to address the U.S. and global needs for meeting energy requirements using clean energy. Some of the benefits outlined include: (i) a plethora of opportunities for different types of industries such as the mining and extraction, manufacturing, and supply chain industries; (ii) CCUS technology in the U.S. may also keep projects that rely on coal and natural gas in 2040 viable in the long run whilst simultaneously reducing carbon emissions.

Lastly, a study conducted by the U.S. Department of the Interior, Bureau of Ocean Energy Management (BOEM) (Vidas, Hugman, Chikkatur, & Venkatesh, 2012) investigated the potential for storing CO_2 in the outer continental shelf of the United States. The authors state that the storage of CO_2 in offshore areas of the U.S. provides additional storage capacity to what is already being stored onshore. The storage of carbon offshore may also lower the political, environmental, and societal hurdles that onshore storage sites have to overcome. The authors point out that the offshore storage projects may prove to be easier to site, permit, finance and operate. Moreover, it may prove faster to execute these projects and develop them as compared to their onshore counterparts. The authors conclude that, based on the assumptions they have made in their study, the geological storage projects that are planned in the outer continental shelf of the U.S. may provide an undiscounted cumulative net benefit to the U.S. economy of $16.9 billion between 2015 and 2050. It should be noted however that these benefits are highly dependent on the assumptions that have been made in the study that include (i) the timing and severity of GHG regulations, (ii) degree to which CCS projects are subsidized by the government, and (iii) the cost of onshore geological storage of CO_2.

1.6. Preliminary Analysis and Results

1.6.1. Description of Field

The extract below (Equinor, 2018) describes the research area (the Volve field):

> *"The Volve field was discovered in 1993 (Discovery well 15/9-19 SR). Located in the central part of the North Sea, 5 km north of Sleipner East, waterdepth 80 m.*
>
> *Volve was developed when oil prices were low. A non-traditional concept was chosen to extract resources easily and profitably. The plan for development and production was approved by NPD in 2005. A description of field development can be found in the PUD report.*
>
> *The Volve development took advantage of an opportunity to use a vacant new building, Maersk Inspirer, which at the time was the world's largest jack-up rig. The rig was equipped with its own wellhead and process module for drilling and production. After the shutdown of the Volve field, this rig will be used in the Repsol-operated Yme field which is to be reopened for production.*
>
> *FSU Navion Saga was used as a storage vessel for the oil. Both installations were leased under the license and operation of the contractors Mærsk Drilling and Teekay.*
>
> *The reservoir in the Volve field consists of Jurassic sandstones at 2750 – 3120 m TVD ss, the field was produced with water injection as pressure support.*
>
> *When the Volve field started well drilling in May 2007, and came into production the following year, life expectancy was from 3-5 years. New wells were being drilled up until 2012-13, which contributed to the increased recovery rate and extended life of the field. However remaining resources were very limited and with the decrease in oil price over recent years, new wells were no longer profitable.*
>
> *The Volve oil and gas field closed in 2016, having continued production 3 years longer than originally planned.*
> *All possibilities to extend the life of the field were explored, which yielded very good results.*
>
> *The field was in production for over 8 years and delivered about 9.5 million barrels of oil beyond what was expected in the development and operation plan. In total, Volve achieved a recovery rate of 54%.*
>
> *Volve was shut down in 2016 by decision of the partners: Statoil (now Equinor), ExxonMobil and Bayern Gas (Now Spring Energy), and with the approval of OED. Production ended in October 2016 and FSU Navion Saga left the field. Permanent plugging of the wells and other process activities were carried out in December. Maersk Inspirer then disconnected from the Volve field. Removal of Subsea equipment was completed in the summer of 2017."*

1.6.2. Description of Log Data

Refer to the extract below (Equinor, 2018) for a brief description of the well log data:

> *"Well logs contain a variety of logs: mud logs, petrophysical logs, vsp, core, biostrat, geochem and corresponding reports from 3 exploration wells and 21 production/injection wells. The data is organized by datatype. The folder contains a file "Volve Inventory" with an overview of the data in the folder."*

1.6.3. Description of Core Data

A conventional core analysis was performed for the majority of the wells that were drilled in the Volve Field, North Continental Shelf. The analysis involved conducting a variety of laboratory measurements such as spectral core gamma log, fluid saturation, tracer analysis, gas permeability, porosity and grain density, liquid permeability, and formation resistivity factor. Thin sections were also prepared from the extracted core plugs. For instance, in Well 15/9-19 A a conventional core analysis was conducted on the 14th of April 1998 on 728 core plugs which were extracted from 7 core samples collected at different depth intervals. The following are the core depth intervals for Well 15/9-19 A:

1. 3837.00 – 3852.22 m
2. 3854.00 – 3881.65 m
3. 3881.50 – 3908.40 m
4. 3908.50 – 3934.50 m
5. 3935.50 – 3963.00 m
6. 3963.00 – 3991.00 m
7. 3991.00 – 4016.70 m

The results of the core analysis provided data on gas and liquid permeability in the horizontal and vertical directions, porosity, grain density in the horizontal and vertical directions, and general descriptions on the lithology observed. Photographs of the core sample were also taken at different intervals. ***Figure 1.7*** illustrates an example from Well 15/9-19 A at a depth interval of 3837 – 3842 m.

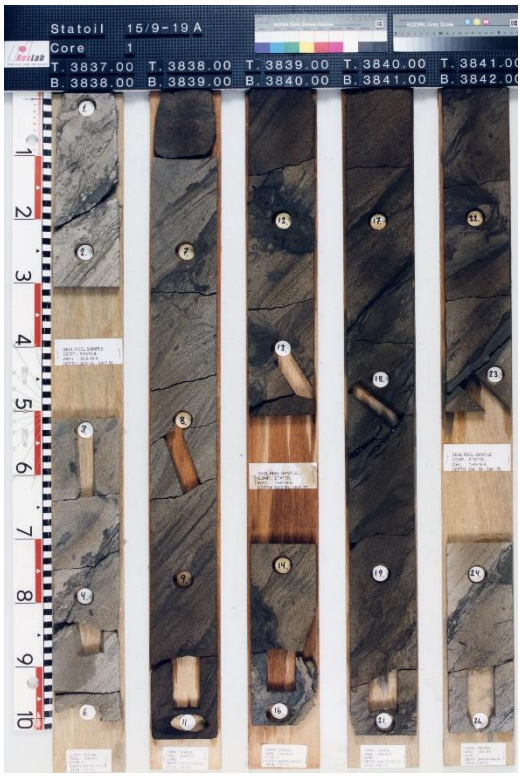


Figure 1.7 Core photograph from Well 15/9-19 A at a depth interval of 3837 - 3842 m.

1.6.4. Description of Seismic Data

Refer to the extract below (Equinor, 2018) for a description of the seismic data.

"Seismic contains seismic data of 2 OBC surveys:

- *ST0202 acquired in 2002*
- *ST10010 acquired in 2010.*

ST0202 and ST10010 surveys have different but related IL/XL system:

- *XL(ST10010) = XL(ST0202)*
- *IL(ST10010) = 2 x IL(ST0202) - 1*

The subdirectory ST0202 contains 3D processed data for ST0202

The subdirectory ST10010 contains 3D processed data for ST10010

The subdirectory ST0202vsST10010_4D contains data from 4D processing of ST0202 and ST10010 (matched data).

Names used in the files indicating processing:

- *ST0202R08: Survey ST0202 re-processed in 2008 (PZ and PS imaging in depth) with 2008 velocity model*
- *ST0202ZC11: Survey ST0202 re-reprocessed in 2011 (PZ imaging in depth), in connection with timelapse processing with 2011 velocity model*

- *ST10010ZC11: Survey ST1010 processed in 2011 (PZ imaging in depth) with the 2011 velocity model.*
- *ST0202ZDC12: Survey ST0202 re-processed in 2012 for 4D purpose*
- *ST10010ZDC12: Survey ST10010 re-processed in 2012 for 4D purpose"*

1.7. Conclusion

To conclude, it is hoped that the work that is proposed in this thesis proposal will be able to provide practitioners with a general workflow on how to locate suitable CO_2 sequestration sites by making use of the massive amount of data already available from the Upstream Oil & Gas Industry alongside the use of artificial intelligence techniques such as artificial neural networks and fuzzy logic.

1.8. References

Abdulraheem, A., Sabakhi, E., Ahmed, M., Vantala, A., Raharja, I., & Korvin, G. (2007). Estimation of Permeability from Wireline Logs in a Middle Eastern Carbonate Reservoir using Fuzzy Logic. *15th SPE Middle East Oil & Gas Show and Conference* (pp. 1-11). Bahrain International Exhibition Centre, Kingdom of Bahrain: Society of Petroleum Engineers.

Ahn, S., Park, C., Kim, J., & Kang, J. M. (2018). Data-Driven Inverse Modeling with a Pre-Trained Neural Network at Heterogeneous Channel Reservoirs. *Journal of Petroleum Science and Engineering*. doi:10.1016/j.petrol.2018.06.084

Altintas, I. (2018). Introduction to Big Data. *The Hadoop Ecosystem: Welcome to the zoo!* San Diego, California, United States of America: Coursera.

Aminzadeh, F. (1994). Applications of Fuzzy Experts Systems in Integrated Oil Exploration. *Computers Elect. Engng*, *20*(2), 89-97.

Aminzadeh, F., & Brouwer, F. (2006). Integrating Neural Networks and Fuzzy Logic for Improved Reservoir Property Prediction and Prospect Ranking. *SEG/New Orleans 2006 Annual Meeting* (pp. 1752-1756). New Orleans: Society of Exploration Geophysicists.

Anifowose, F. A., Abdulraheem, A., Al-Shuhail, A., & Schmitt, D. P. (2013). Improved Permeability Prediction from Seismic and Log Data using Artificial Intelligence Techniques. *SPE Middle East Oil and Gas Show and Conference* (pp. 1-7). Manama, Bahrain: Society of Petroleum Engineers.

Baaziz, A., & Quoniam, L. (2014). How to use Big Data technologies to optimize operations in Upstream Petroleum Industry. *21st World Petroleum Congress* (pp. 1-9). Moscow: 21st World Petroleum Congress.

Bachu, S. (2000). Sequestration of CO2 in Geological Media: Criteria and Approach for Site Selection in Response to Climate Change.

Energy Conversion & Management, 953-970.

Bachu, S. (2002). Sequestration of CO2 in Geological Media in Response to Climate Change: Road Map for Site Selection using the Transform of the Geological Space into the CO2 Phase Space. *Energy Conversion & Management*, 87-102.

Baines, S. J., & Worden, R. H. (2004). Geological storage of carbon dioxide. *Geological Society, London, Special Publications*, 1-6.

Benson, S. M., & Franklin M. Orr, J. (2008). Carbon Dioxide Capture and Storage. *MRS Bulletin*, 303-305. doi:10.1557/mrs2008.63

Chadwick, R., Zweigel, P., Gregersen, U., Kirby, G., Holloway, S., & Johannessen, P. (2004). Geological Reservoir Characterization of a CO2 Storage Site: The Utsira Sand, Sleipner, Northern North Sea. *Energy*, 1371-1381.

Cuddy, S. J. (1997). The Applications of the Mathematics of Fuzzy Logic to Petrophysics. *SPWLA 38th Annual Logging Symposium* (pp. 1-14). Houston, Texas: Society of Petrophysicists and Well Log Analysts.

Cuddy, S. J. (2000). Litho-Facies and Permeability Prediction from Electrical Logs using Fuzzy Logic. *SPE Reservoir Eval. & Eng., 3*(4), 319-324.

Equinor. (2018, May 21). *Volve data village*. Retrieved July 26, 2018, from Equinor: https://data-equinor-com.azurewebsites.net/dataset/volve

Esmaili, S., & Mohaghegh, S. D. (2016). Full Field Reservoir Modeling of Shale Assets using Advanced Data-Driven Analytics. *Geoscience Frontiers*, 11-20.

Fang, Y., Baojun, B., Dazhen, T., Dunn-Norman, S., & Wronkiewicz, D. (2010). Characteristics of CO2 Sequestration in Saline Aquifers. *Petroleum Science*, 83-92.

Firican, G. (2018). *The 10 Vs of Big Data*. Retrieved from TDWI: https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Gale, J. (2004). Geological storage of CO2: What do we know, where are the gaps and what more needs to be done? *Energy*, 1329-1338.

Hardisty, P. E., Sivapalan, M., & Brooks, P. (2011). The Environmental Economic Sustainability of Carbon Capture and Storage. *International Journal of Environmental Research and Public Health*, 1460-1477.

Haris, A., Fakhri, M. K., Isniarny, N., & Riyanto, A. (2017). A Case Study of Hydrocarbon Prospect Evaluation in North Exito Field by using Seismic Attribute and Seismic Inversion. *American Institute of Physics*, (pp. 1-4).

Harrison, B., & Kennedy, M. (2002). Improving Prospect Evaluation by Integrating Petrophysical Models into the Workflow. *SCA*, (pp. 1-12).

Hassanzadeh, H., Pooladi-Darvish, M., & Keith, D. W. (2007). Scaling Behavior of Convective Mixing, with Application to Geological Storage of CO2. *AIChe Journal*, 1121-1131.

HelpSaveNature. (2017, December 18). *Role of Carbon Sequestration and its Pros and Cons You Never Knew*. Retrieved September 27, 2018, from HelpSaveNature: https://helpsavenature.com/explanation-of-carbon-sequestration-with-pros-cons

Hui, Y., Baihong, W., Yan, Z., Guangya, Z., Zhizhou, L., Fengcheng, W., . . . Qinghui, H. (2009). Distribution of Hydrocarbon Traps in Volcanic Rocks and Optimization for Selecting Exploration Prospects and Targets in Junggar Basin: Case Study in Ludong-Wucaiwan Area, NW China. *Petroleum Exploration and Development*, 419-427.

Intergovernmental Panel on Climate Change. (2005). *Special Report on Carbon Dioxide Capture and Storage.* Cambridge, UK: Cambridge University Press.

InternetLiveStats. (2018). *Google Search Statistics*. Retrieved July 24, 2018, from InternetLiveStats: http://www.internetlivestats.com/google-search-statistics/

Kaszuba, J. P., Janecky, D. R., & Snow, M. G. (2003). Carbon dioxide reaction processes in a model brine aquifer at 200 °C and 200 bars: implications for geologic sequestration of carbon. *Applied Geochemistry*, 1065-1080.

Kovscek, A. R. (2002). Screening Criteria for CO2 Storage in Oil Reservoirs. *Petroleum Science and Technology*, 841-866.

Lim, J.-S., & Kim, J. (2004). Reservoir Porosity and Permeability Estimation from Well Logs using Fuzzy Logic and Neural Networks. *SPE Asia Pacific Oil and Gas Conference and Exhibition* (pp. 1-9). Perth, Australia: Society of Petroleum Engineers.

Malvić, T., Velić, J., Horváth, J., & Cvetković, M. (2010). Neural Networks in Petroleum Geology as Interpretation Tools. *Central European Geology*, 97-115.

Marr, B. (2015). *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance.* Cornwall: John Wiley & Sons Ltd.

MongoDB, Inc. (2018). *What is MongoDB*. Retrieved from MongoDB: https://www.mongodb.com/what-is-mongodb

Montgomery, D. C., & Runger, G. C. (2014). *Applied Statistics and Probability for Engineers* (6th ed.). Singapore: John Wiley & Sons Singapore Pte. Ltd.

Nashawi, I., & Malallah, A. (2010). Permeability Prediction from Wireline Logs using Fuzzy Logic and Discriminant Analysis. *SPE Asia Pacific Oil & Gas Conference and Exhibition* (pp. 1-12). Brisbane, Queensland, Australia: Society of Petroleum Engineers.

Olatunji, S., Selamat, A., & Abdulazeez, A. (2015). Harnessing the Power of Type-2 Fuzzy Logic System in the Prediction of Reservoir Properties. *SPE Saudi Arabia Section Technical Symposium and Exhibition* (pp. 1-24). Al-Khobar, Saudi Arabia: Society of Petroleum Engineers.

Otis, R. M., & Schneidermann, N. (1997). A Process for Evaluating Exploration Prospects. *AAPG Bulletin*, 1087-1109.

Oyeyemi, K. D., Olowokere, M. T., & Aizebeokhai, A. P. (2018). Hydrocarbon Resource Evaluation using Combined Petrophysical Analysis and Seismically Derived Reservoir Characterization, Offshore Niger Delta. *J Petrol Explor Prod Technol*, 99-115.

Patrizio, A. (2018, May 30). *Big Data vs. Artificial Intelligence*. Retrieved from Datamation: https://www.datamation.com/big-data/big-data-vs.-artificial-intelligence.html

Qmee. (2013). *Online in 60 Seconds*. Retrieved July 24, 2018, from Qmee: https://blog.qmee.com/wp-content/uploads/2013/07/Qmee-Online-In-60-Seconds2.png

Raef, A., Meek, T. N., & Totten, M. W. (2016). Applications of 3D Seismic Attribute Analysis in Hydrocarbon Prospect Identification and Evaluation: Verification and Validation based on Fluvial Paleochannel Cross-Sectional Geometry and Sinuosity, Ness Country, Kansas, USA. *Marine and Petroleum Geology*.

Roisenberg, M., Schoeninger, C., & Silva, R. R. (2009). A Hybrid Fuzzy-Probabilistic System for Risk Analysis in Petroleum Exploration Prospects. *Expert Systems with Applications*, 6282-6294.

Saggaf, M. M. (2002). Estimation of Reservoir Quality by Attribute Integration through Fuzzy Logic. *17th World Petroleum Congress* (pp. 231-241). Rio de Janeiro, Brazil: Society of Petroleum Engineers.

Spofforth, D., & Firth, J. (2016). Integrated Datasets Hold the Key to Unravelling Petroleum Prospectivity. *World Oil® Special Focus: Advances in Exploration*, 33-38.

Taghavi, A. (2005). Improved Permeability Estimation through use of Fuzzy Logic in a Carbonate Reservoir from Southwest Iran. *14th SPE Middle East Oil & Gas Show and Conference* (pp. 1-9). Bahrain International Exhibition Centre, Bahrain: Society of Petroleum Engineers.

The Apache Software Foundation. (2008). *HDFS Architecture Guide*. Retrieved from Hadoop: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

The Apache Software Foundation. (2013). *MapReduce Tutorial*. Retrieved from Hadoop: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

The Apache Software Foundation. (2014). *Apache Hive*. Retrieved from Apache Hive: http://hive.apache.org/

The Apache Software Foundation. (2015). *Apache Storm*. Retrieved from Apache Storm: http://storm.apache.org/index.html

The Apache Software Foundation. (2016). *What is Cassandra*. Retrieved from Apache Cassandra: http://cassandra.apache.org/

The Apache Software Foundation. (2017). *Welcome to Apache Zookeeper™*. Retrieved from Apache ZooKeeper™: http://zookeeper.apache.org/

The Apache Software Foundation. (2017). *What is Apache Flink?* Retrieved from Apache Flink: https://flink.apache.org/flink-architecture.html

The Apache Software Foundation. (2018). *Apache Hadoop YARN*. Retrieved from Apache Hadoop: https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

The Apache Software Foundation. (2018). *Welcome to Apache Giraph!* Retrieved from Apache Giraph: http://giraph.apache.org/

The Apache Software Foundation. (2018). *Welcome to Apache HBase™*. Retrieved from Apache HBase: http://hbase.apache.org/

The Apache Software Foundation. (2018). *Welcome to Apache Pig!* Retrieved from Apache Hadoop: http://pig.apache.org/

The Apache Software Foundation. (2018). *Welcome to Apache™ Hadoop®!* Retrieved from Apache Hadoop: http://hadoop.apache.org/

The MathWorks, Inc. (2009). *Neural Network Toolbox™ 6 User's Guide.* The MathWorks, Inc.

The MathWorks, Inc. (2010). *Fuzzy Logic Toolbox™ 2 User's Guide.* The MathWorks, Inc.

Torp, T. A., & Gale, J. (2004). Demonstrating storage of CO2 in geological reservoirs: The

Sleipner and SACS projects. *Energy*, 1361-1369.

U.S. Department of Energy. (2016). *Carbon Capture, Utilization, and Storage: Climate Change, Economic Competitiveness, and Energy Security.* U.S. Department of Energy.

University of Birmingham. (2018). *How to Write a Research Proposal*. Retrieved from University of Birmingham: https://www.birmingham.ac.uk/schools/law/courses/research/research-proposal.aspx

Vidas, H., Hugman, B., Chikkatur, A., & Venkatesh, B. (2012). *Analysis of the Costs and Benefits of CO2 Sequestration on the U.S. Outer Continental Shelf.* U.S. Department of the Interior, Bureau of Ocean Energy Management.

West, J. M., Pearce, J., Bentham, M., & Maul, P. (2005). Issue Profile: Environmental Issues and the Geological Storage of CO2. *European Environment*, 250-259.

Workflows for Data Science Center of Excellence. (2018). *Big Data*. Retrieved July 24, 2018, from Workflows for Data Science Center of Excellence: https://words.sdsc.edu/words-data-science/big-data

Yelebe, Z. R., & Samuel, R. J. (2015). Benefits and Challenges of Implementing Carbon Capture and Sequestration Technology in Nigeria. *The International Journal of Engineering and Science*, 42-49.

2. CHAPTER 2 - PERMEABILITY PREDICTION VIA ARTIFICIAL NEURAL NETWORKS (ANN)

2.1. Abstract

The storage of CO_2 in deep geological formations has been put forth as a solution to tackle excessive anthropogenic CO_2 emissions. This practice leads to the well-established phenomenon of global warming. Several studies have aimed to address the issues pertaining to this proposed solution that range from the costs of building the necessary infrastructure to capture, transport, and store CO_2 in these geological formations. Additionally, the monitoring of the CO_2 for excessive periods of time after the carbon capture and storage (CCS) project has been concluded to ensure no leakage to the surface, has posed several difficulties for large scale implementation.

When planning for a CCS project, several factors come into play with regards to how one would select a suitable storage site. A review of literature in this research area reveals that the process of characterizing a suitable site in terms of geological aspects is similar to how the Oil & Gas exploration industry would go about characterizing a potential hydrocarbon prospect. This presents a major advantage to the storage of CO_2, in the sense that the massive amounts of data (well log data, seismic data, and core data) that would be necessary for a proper site characterization would already be present from the exploration of hydrocarbon reserves in a prospective area.

In this paper, we focus on the use of well log data obtained from a well in the Volve field present in the Norwegian Continental Shelf to predict permeability data. The output is compared to permeability data obtained from a routine analysis of 7 different core samples collected from the same well. Three well logs were used; gamma ray (GR), bulk density (RHOB), and neutron porosity (NPHI), to predict permeability data that was obtained from 557 core plugs. To this end, a single layered artificial neural network (ANN) was utilized to train 70% of the data points and test the remaining 30% of the data. The optimization of results was performed by comparing three main statistical parameters, which are the correlation coefficient (CC), the root mean-squared error (RMSE) and the absolute average percentage error (AAPE).

2.2. Introduction

CO_2 sequestration has been put forth as one of the solutions to the problem of global warming and climate change in general. These issues are increasingly becoming more urgent and the window for acting upon them is narrowing. The process of CO_2 sequestration requires CO_2 to first be captured from the atmosphere (Benson & Franklin M. Orr, 2008), or other sources that emit CO_2. The CO_2 is then transported to where it can be injected, via an injection well, to a geological formation. The geological formation should be such that the CO_2 remains trapped for thousands of years. There are several options as regards to what type of geological formation may be used for this purpose (Intergovernmental Panel on Climate Change, 2005). Attempts have been made at storing it in (i) depleted oil and gas reservoirs, (ii) coal seams that are no longer mineable, and (iii) deep saline water aquifers.

The question that may arise is how would one go about the selection of a prospective CO_2 sequestration site? Literature review in this subject area reveals that there are number of factors that would need to be taken into account in order to select a potential site. These factors may include **basin** and **reservoir** characteristics. Examples are: the type of tectonic setting that the reservoir was formed in, the geothermal regime in the area as the temperature of the subsurface would influence the density of the injected CO_2, the presence of geologic structures such as anticlines and salt domes that may serve as zones of accumulation for injected CO_2, the strength and integrity of the sealing cap rock layer that extends above a reservoir, and general reservoir properties such as porosity and permeability that give an indication about the ability of the reservoir to store and transmit fluids contained in the pore space (Fang, Baojun, Dazhen, Dunn-Norman, & Wronkiewicz, 2010). Secondly, the presence of discontinuous local sealing layers that may serve as barriers to the migration of CO_2 to the surface (e.g. tight sandstone stringers or shale layers that are characterized by permeability values on the scale of nanoDarcies), the existence of structural and stratigraphic traps throughout a prospective area, the structure and distribution of geological layers of the reservoir itself, and the heterogeneity of the reservoir (Chadwick, et al., 2004). Thirdly, in case of sequestration into an oil reservoir, one may need to look at **reservoir engineering aspects** such as the density of the injected CO_2 (as previously mentioned, this may be influenced by the geothermal regime), the ability of or the extent to which the reservoir oil can saturate itself with CO_2, the interaction of a bordering water aquifer with the oil reservoir, and the amount of incremental recovery that may be attained via the injection of the CO_2 (in this particular case, for enhanced oil recovery (EOR) purposes); **geophysical aspects** such as the presence of seals, natural small-scale and large-scale fractures, structures and favorable stratigraphy; and lastly, **surface facilities aspects** such as the availability of the necessary technology and infrastructure to compress, transport, store, and transport the CO_2 to the selected injection site, and the capability to handle any corrosion that may be caused by the generally impure CO_2 gas stream that may contain corrosive impurities such as H_2S (hydrogen sulfide) (Kovscek, 2002). Furthermore, the hydrodynamic regimes of water located in the pore space, potential of the hydrocarbon present, maturity of the basin underlying the reservoir, economic aspects related to building the necessary infrastructure, transporting the gas stream, compressing and storing it if required, and injecting it into the formation, and the political and social conditions as regards to the large-scale implementation of this technology need to be taken care of (Bachu, 2000). Lastly, the **basin geology and hydrostratigraphy** needs to be understood in order to identify areas that are characterized by confinement that would serve to trap the CO_2 gas, and avoid areas that may pose potential migration pathways for the CO_2 to seep to the surface; **basin hydrodynamics** can also help identify areas that may have a potential to trap the CO_2 hydrodynamically (e.g. locations of regional recharge or erosional rebound) (Bachu, 2002).

As can be seen from the discussion above, there are numerous factors that need to be taken care of for a full-scale screening process to be conducted in order to select an ideal CO_2 sequestration site. For the purposes of this work, however, the discussion is limited to geological and geophysical aspects.

The discussion will further be narrowed down to the determination of favorable permeability that would allow the injected CO_2 gas stream to travel to the determined zones of accumulation with ease via the use of *Artificial Neural Networks* (ANN).

There has been a great deal of studies performed in the upstream Oil & Gas Industry on the utilization of artificial intelligence techniques to determine reservoir properties such as formation permeability. Abdulraheem *et al*. (2007) made use of fuzzy logic to predict permeability by the use of conventional open-hole logs. The fuzzy logic algorithm was implemented by the use of MATLAB, where the model was trained on 70% of the data and the remaining 30% of the data were used for testing the generated model. The authors also attempted to predict permeability for new layers or wells. The authors concluded that the use of fuzzy logic produces good results in predicting permeability. Olatunji *et al*. (2015) attempted to improve uncertainty handling in the prediction of permeability and PVT properties by making use of a type-2 fuzzy logic system. To achieve this, the authors developed a hybrid system that would first pass their dataset through the type-2 fuzzy logic system. The output coming out of this first stage would then be used to train another portion of the hybrid scheme. 70% of the data set was used to train the data, and the remaining 30% was tested by the use of the developed hybrid system. The authors claim that this type-2 fuzzy logic system offers an improved performance over individual models, in the case of permeability and PVT property prediction. Taghavi (2005) explored the performance of three different methods for predicting permeability which included (i) relating the permeability with porosity values obtained from core plugs, (ii) use of a multilinear regression, and (iii) use of fuzzy logic. The other compared these different methods and concludes that fuzzy logic out-performed the previous two in the prediction of permeability. Anifowose *et al*. (2013) showed that combining seismic data along with log data can enhance the prediction of permeability. The authors compared the performance of using log data alone, seismic data alone, and the combination of the two to predict formation permeability using three different AI techniques (ANN, SVM, Fuzzy Logic). Cuddy (1997, 2000) utilized fuzzy logic to predict litho-facies and permeability. In the case of litho-facies prediction, the author used descriptions from 10 cored wells to predict the facies in 30 uncored wells, whereas in the case of permeability prediction the author used core permeability values obtained from one well to train the model, the model was then used to predict permeability in another well. Nashawi and Malallah (2010) proposed a two-step process to predict formation permeability from wireline log data. The process involves using data mining algorithms to first classify different types of rocks, and then applying experiment system techniques to predict the output. The analysis that they performed included principle component analysis, cluster analysis, and discriminant analysis. Lastly, Lim and Kim (2004) made use of fuzzy logic as well as ANN to determine reservoir properties. The authors used fuzzy logic to select the best related cores with well logs and core porosity and permeability data. This offers an alternative to the generally used statistical parameters (such as correlation coefficient) to select the best input data to use. They then made use of a neural network to develop the transformation between the log and core data. The model is then validated to well data in a well located in offshore Korea.

The results of the prediction of formation permeability from these references have been summarized in ***Table 2.1***. Unfortunately, the majority of these listed references have not explicitly reported the results of their prediction. This makes it difficult to gauge the performance of their predictions with those of other investigators in the literature.

Table 2.1 Summary of statistical parameters reported from listed references. Parameters include the coefficient of determination (R^2), the absolute average percentage error (AAPE), and the root mean squared error (RMSE). As can be seen from the table, the majority of the authors have not explicitly reported these results.

Reference	*Coefficient of Determination (R^2)*	*Absolute Average Percentage Error (AAPE)*	*Root Mean Squared Error (RMSE)*
(Abdulraheem, et al., 2007)	N/A	N/A	N/A
(Olatunji, Selamat, & Abdulazeez, 2015)	0.9543	70.3821	0.2735
(Taghavi, 2005)	N/A	N/A	N/A
(Anifowose, Abdulraheem, Al-Shuhail, & Schmitt, 2013)	CC: ~0.425	N/A	~1.20
(Cuddy, 1997)	N/A	N/A	N/A
(Cuddy, 2000)	N/A	N/A	N/A
(Nashawi & Malallah, 2010)	N/A	N/A	N/A
(Lim & Kim, 2004)	N/A	N/A	N/A

ANN has been used quite extensively over the past few decades in the prediction of formation properties. In this section we briefly look at the more recent developments in using ANN to predict formation permeability. ***Table 2.2*** summarizes the results of the prediction from these recent studies.

Jamshidian *et al*. (2015) made use of an ANN coupled with a multi-linear perception (MLP) structure and feed-forward back-propagation algorithm to attempt to predict the parameters associated with the nuclear magnetic resonance (NMR) log. The authors utilized a hybrid type model, where the parameters of the ANN were optimized by use of an imperialist competitive algorithm (ICA). Data obtained from one well were used to train and develop the model, and the model was then tested with data from another well. The authors conclude that the ANN-ICA performs well in the prediction of permeability and free flowing porosity. Saljooghi and Hezarkhani (2015) investigated the performance of a wavenet in the prediction of permeability from well log data and compared the results to those of an ANN. They utilized 190 samples for training purposes, tested the model on 35 samples, and validated with 35 samples. The authors conclude that the wavenet produced better results than the ANN in terms of the correlation coefficient obtained between the predicted and actual results. Iturrarán-Viveros and Parra (2014) made use of an ANN model to predict permeability and porosity from a carbonate reservoir present in southeastern Florida. Moreover, the also used the model to predict intrinsic attenuation for a sand-shale oil reservoir present in northeast Texas. The authors report the performance of the model in the prediction of the intrinsic attenuation and conclude that the ANNs obtained are capable of providing reliable and realistic predictions for the three outputs being investigated. Bagheripour (2014) developed a committee neural network (CNN) to predict rock permeability. The CNN was developed by the use of a genetic algorithm to

integrate different ANNs that included multi-layer perceptron (MLP), radial basis function (RBF), and generalized regression neural network (GRNN). To develop the model the authors used 350 data points to establish the CNN and 245 data points to test its reliability. Rafik and Kamel (2017) attempted to predict the permeability for a reservoir present in Algeria. They made use of three non-parametric approaches that included alternating conditional expectations (ACE, generalized additive model (GAM), and neural networks (NNET). To develop and test the models the authors utilized 927 sample data points from 7 wells. Finally, Elkatatny *et al*. (2018) developed an ANN that could prediction formation permeability with the use of just three well logs. Model development and testing was conducted by training 70% of the data points (857 data points) and testing with the remaining 30% (366 data points). The authors compared the performance of the developed ANN model with ANFIS and SVM. They conclude that the ANN is preferable since a mathematical expression to calculate permeability may be extracted from the network.

Table 2.2 Summary of statistical parameters reported from recent studies on using ANN to predict formation permeability. Parameters include the coefficient of determination (R^2), the absolute average percentage error (AAPE), and the root mean squared error (RMSE).

Reference	*Case/AI Technique*	*Coefficient of Determination (R^2)*	*Absolute Average Percentage Error (AAPE)*	*Root Mean Squared Error (RMSE)*
(Jamshidian, et al., 2015)	*Without Optimization*	0.8567	N/A	MSE: 0.1174
	With Optimization	0.8975	N/A	MSE: 0.0303
(Saljooghi & Hezarkhani, 2015)	*ANN*	0.86	N/A	0.39
	Wavenet	0.90	N/A	0.37
(Bagheripour, 2014)	*MLP*	0.8141	N/A	0.0285
	GRNN	0.8092	N/A	0.0295
	RBF	0.8311	N/A	0.0263
	CNN	0.8331	N/A	0.0250
(Elkatatny, Mahmoud, Tariq, & Abdulraheem, 2018)	*ANN*	CC: 0.95	N/A	0.37
	ANFIS	CC: 0.96	N/A	0.37
	SVM	CC: 0.96	N/A	0.38

The fundamental principle of the ANN revolves around the following terms (The MathWorks, Inc., 2009):

- Scalar or vector input
- Weight
- Bias
- Transfer function
- Output

The input is the data that you input into the network. Keep in mind that the quality control (QC) and statistical analysis should have already been performed on this data before uploading it onto the network. The QC of the data will help us determine any anomalies in the data or unexpected trends. The input is then multiplied by a *weight* which is a scalar

quantity. Additionally, the product of the input and the weight is added to a *bias*, which can be viewed as a shift of the *transfer function* that performs the final computation to produce the output. Refer to ***Figure 2.1*** for an illustration of how these terms interact with each other.

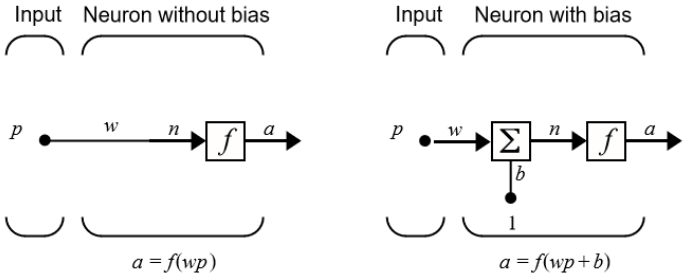


Figure 2.1 The interaction of the different terms used in describing ANN (The MathWorks, Inc., 2009). Where *p* is the input, *w* is the weight, *wp* is the product of the weight and input, *b* is the bias, and *a* is the output.

The transfer function is what is used to convert the modified input to the final output. There are several different transfer functions that are available for use. The transfer function that is best to use would largely depend on the type of problem you are dealing with. Examples include:

- Linear transfer function
- Tan-sigmoid transfer function
- Log-sigmoid transfer function

The *linear transfer function* calculates the neuron's output by simply returning the value passed to it. In other words, it produces its input from its output: $a = purelin(n) = purelin(Wp + b) = Wp + b$.

This neuron can be trained to learn an affine function of its inputs, or to find a linear approximation to a nonlinear function. A linear network cannot, of course, be made to perform nonlinear computations (see ***Figure 2.2***).

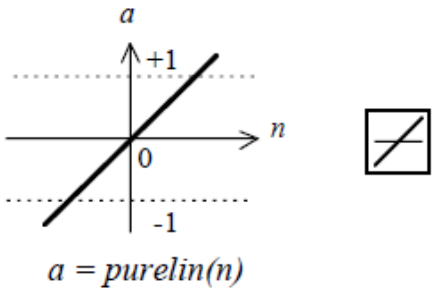


Figure 2.2 Linear Transfer Function (The MathWorks, Inc., 2009).

The *log-sigmoid transfer function* is a function which compresses or 'squashes' the input that is provided to it to range in between 0 and 1. The function is given by the following equation: $f(n) = \frac{1}{1+e^{-n}}$ (see ***Figure 2.3***).

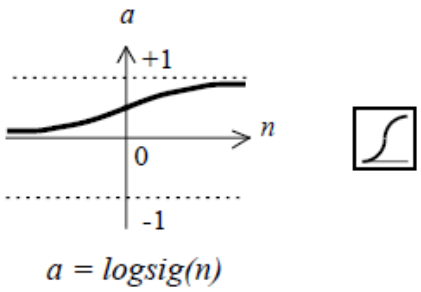


Figure 2.3 Log-Sigmoid Transfer Function (The MathWorks, Inc., 2009).

Similarly, the *tan-sigmoid transfer function* is a function which compresses or 'squashes' the input that is provided to it to range in between -1 and 1. The function is given by the following equation: $f(n) = \frac{1}{1+e^{-n}}$ (see ***Figure 2.4***).

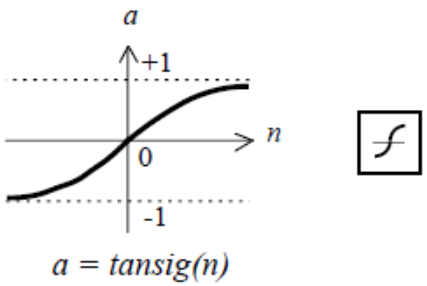


Figure 2.4 Tan-Sigmoid Transfer Function (The MathWorks, Inc., 2009).

A simple approach would be using a variety of these to determine which will produce the best results.

In this paper, we focus on the use of well log data obtained from a well in the Volve field present in the Norwegian Continental Shelf to predict permeability data obtained from a routine core analysis of 7 different core samples obtained from the same well. Three well logs were used, that included gamma ray (GR), bulk density (RHOB), and neutron porosity (NPHI), to predict permeability data that was obtained from 557 core plugs.

2.3. Methodology

2.3.1. Data Set

The data used for this study was obtained from *Equinor*'s latest massive data release to the public for the purposes of facilitating research for students and scientists around the world. The data of this release includes geophysical interpretations, production data, well reports, reservoir models, seismic data, well log data, well technical data, and well real-time drilling data. Detailed core data are also provided for several wells. This data was collected from the *Volve* field, located in the Norwegian continental shelf. The Volve field was first discovered in 1993 and was put into production in 2008. Production ceased in 2016 and the field was then planned for decommissioning by the end of 2018. Volve is located north of the *Sleipner Øst* field in the central part of the North Sea at a distance of 5 km. The

reservoir section of this field is located at a depth interval of 2750 to 3120 m (Equinor, 2018).

Data from a total of 24 wells are provided, which include mud logs, logging while drilling (LWD) logs, composite logs, petrophysical interpretations, lithology, fluid, and porosity (LFP) information, velocity checkshot data, core data, biostratigraphy data, geochemical logs, and drilling reports. For the purposes of this work, LWD data and core data from one exploratory well (Well 15/9-19 A) will be utilized to train and test the ANN model. Three different and commonly run logs were utilized to conduct the prediction. These are the gamma ray (GR), bulk density (RHOB), and neutron porosity (NPHI) logs. Refer to ***Figure 2.5*** for a snip of the main log head that was obtained from Well 15/9-19 A.

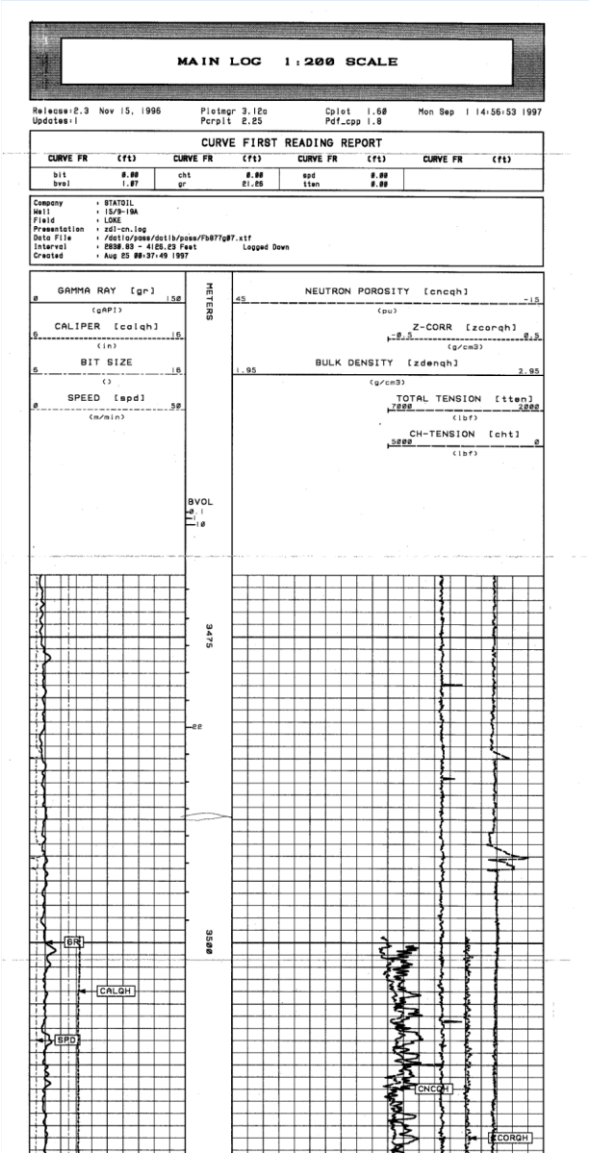


Figure 2.5 Snip of the main log head that was obtained from Well 15/9-19 A. Information included here are the gamma ray, caliper, neutron porosity, and bulk density logs. The total logged interval is from 2838.83 ft to 4126.23 ft.

Conventional core analysis was performed for majority of wells that were drilled in the Volve Field, North Continental Shelf. The analysis involved conducting a variety of laboratory measurements such as spectral core gamma log, fluid saturation, tracer analysis, gas permeability, porosity and grain density, liquid permeability, and formation resistivity factor. Thin sections were also prepared from the extracted core plugs. For instance, in Well 15/9-19 A, a conventional core analysis was conducted on 14th of April 1998 on 728 core plugs which were extracted from 7 core samples collected at different depth intervals. The following are the core depth intervals for Well 15/9-19 A:

1. 3837.00 – 3852.22 m
2. 3854.00 – 3881.65 m
3. 3881.50 – 3908.40 m
4. 3908.50 – 3934.50 m
5. 3935.50 – 3963.00 m
6. 3963.00 – 3991.00 m
7. 3991.00 – 4016.70 m

The results of core analysis provided data on gas and liquid permeability in the horizontal and vertical directions, porosity, grain density in the horizontal and vertical directions, and general descriptions of the lithology observed. Photographs of the core sample were also taken at different intervals. ***Figure 2.6*** illustrates an example from Well 15/9-19 A at a depth interval of 3837 to 3842 m.

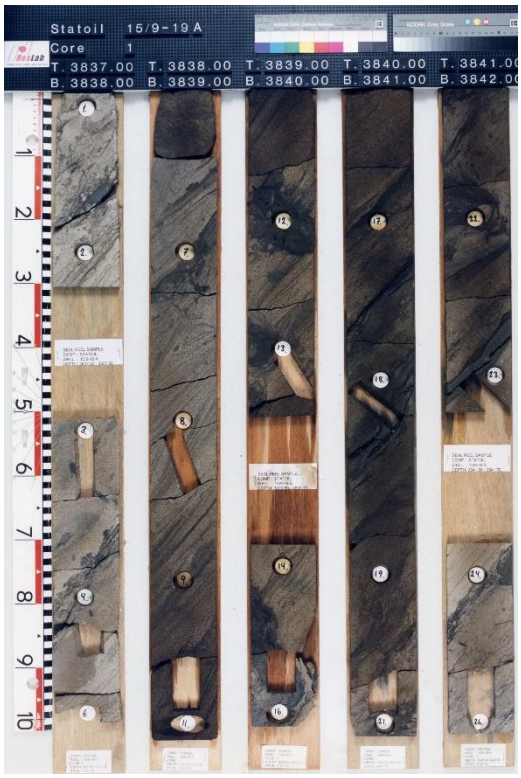


Figure 2.6 Core photograph from Well 15/9-19 A at a depth interval of 3837 to 3842 m.

Well 15/9-19 A is a side track well that was drilled from Well 15/9-19 SR. It was drilled for the purpose of confirming the minimum economic hydrocarbon volume in the Hugin formation as well as to map the overall extension of the oil-bearing formation. Refer to ***Table 2.3*** for additional general information about the well. More information regarding its operations and results obtained from the *Norwegian Petroleum Directorate* (Norwegian Petroleum Directorate, 2018) include the following:

> "*Well 15/9-19 A was kicked off from 2178 m in well bore 15/9-19 SR on 25 July 1997, using the semi-submersible installation Byford Dolphin. The well was drilled through the Skagerrak Formation and terminated approximately 30 m TVD into the Triassic Smith Bank Formation at 4131 m (3318.5 m TVD RKB). The final acquisition programme immediately after reaching the total depth of the well was strongly affected by a labour conflict, which delayed the well operations for 32.5 days. The originally planned open hole electric logging program had to be terminated and the 7" casing run to TD in order to secure the well. The later cased hole logging failed due to tool problems. The well was drilled oil based with the Ultidril mud system (oil base consists of synthetic olefins) from kick-off to TD. Top of the Hugin Formation was penetrated at 3796.5 m (3015.5 m TVD RKB) approximately 60 m TVD deeper than prognosed. It was 153 m thick (TVD) and oil-bearing. The total oil column in the well was 80 m, but no clear oil-water contact was observed. The base of the reservoir was at 3919 m (-3126.5 m TVD RKB). Seven cores were cut in the interval 3838 m to 4017 m in the Hugin and Skagerrak Formations, with a total recovery of 177.6 m. One attempt was made to run FMT on PCL for pressure points and fluid sampling. The run failed for technical reasons and no further attempts were made due to the labour conflict. The well was permanently abandoned on 9 November 1997 as an oil appraisal."*

Table 2.3 General information about Well 15/9-19 A (Norwegian Petroleum Directorate, 2018).

Wellbore Name	15/9-19 A
Type	Exploration
Purpose	Appraisal
Status	Plugged and Abandoned
Main Area	North Sea
Field	Volve
Drilling Operator	Den norske stats oljeselskap a.s
Drill Permit	898-L
Drilling Facility	Byford Dolphin
Drilling Days	108
Entered Date	25.07.1997
Completed Date	09.11.1997
Release Date	09.11.1999
Publication Date	15.12.2006
Kelly Bushing Elevation [m]	25.0
Water Depth [m]	85.0
Total Depth (MD) [m RKB]	4131.0
Final Vertical Depth (TVD) [m RKB]	3319.0
Maximum Inclination [°]	59

Bottom Hole Temperature [°C]	117
Geodetic datum	ED50
NS degrees	58° 26' 9.25" N
EW degrees	1° 55' 47.05" E
NS UTM [m]	6477887.72
EW UTM [m]	437506.71

2.3.2. Research Approach

As can be seen from ***Figure 2.5***, the log data of interest for this well was provided in the form of an image (.TIF file format) as opposed to the typical .LAS or .DLIS file formats that are commonly used for wireline log data. In order to utilize this data, it would have to be digitized first. This was accomplished by use of a MATLAB® code available from MathWork®'s online file exchange (Doke, 2016), which allows one to manually extract data points off an image file. ***Figure 2.7*** shows the digitized input logs. The GR log was calibrated from 0 to 150 gAPI and 3470 to 4127 m. The RHOB log was calibrated from 1.95 to 2.95 g/cc and 3470 to 4127 m. The NPHI log was calibrated from 45 to -14 pu and 3470 to 4127 m. Care needs to be taken to ensure that the peaks of logs are captured.

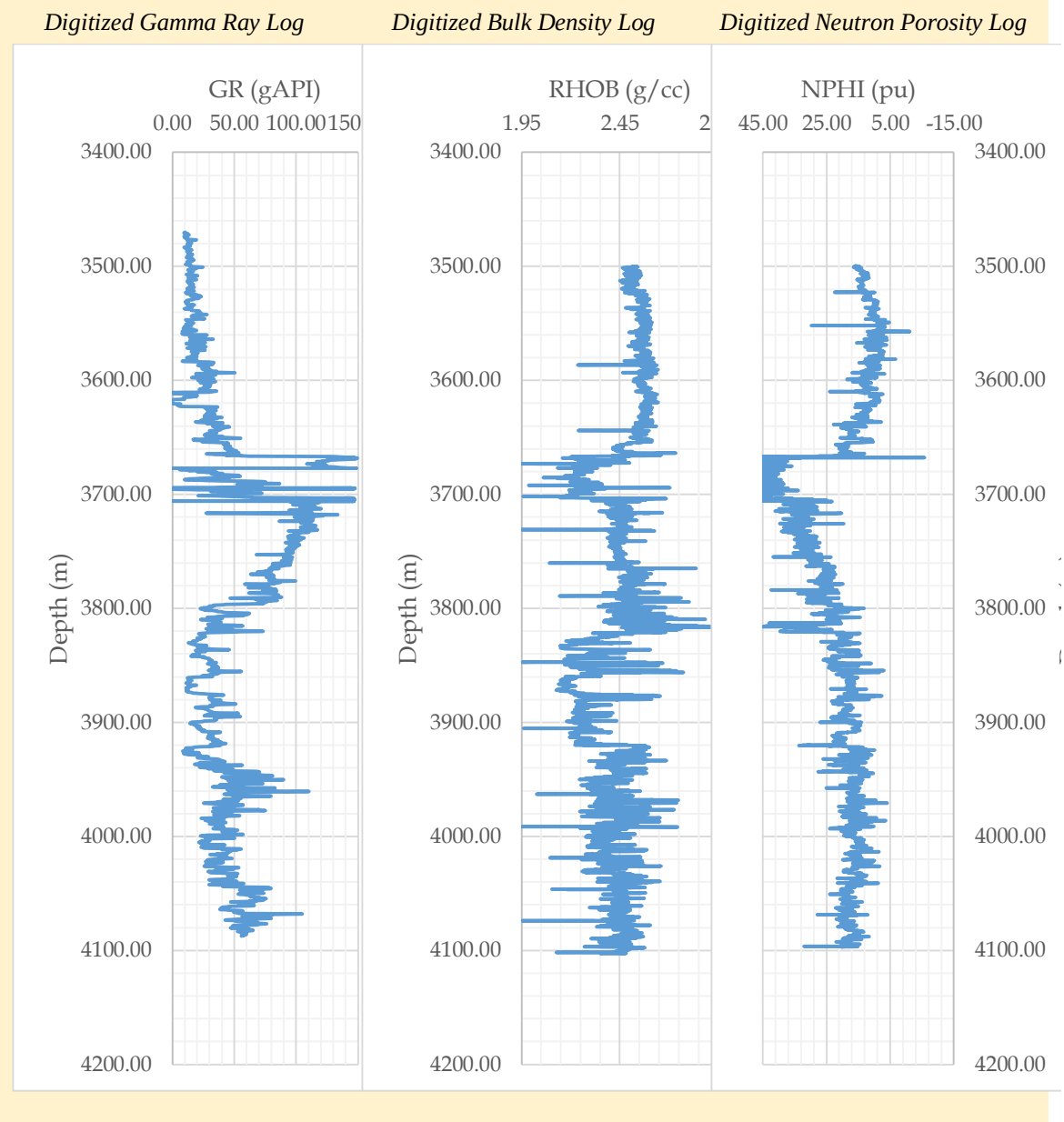


Figure 2.7 Digitized input well logs [Generated data points: GR = 4356 data points, RHOB = 6912 data points, NPHI = 5595 data points].

Once all three input well logs have been digitized, the values of well log at the corresponding core depth, where the permeability measurements were recorded from the extracted core plugs, need to be determined. This can be done by use of a spreadsheet to construct horizontal lines at each of the core depth values. These logs are then enlarged and exported as images to the same MATLAB® code (Doke, 2016) in order to digitize the log values at the intersection point between the log and the constructed horizontal lines. The corresponding number of data points for each of the logs at these core depths equals to 728 - the total number of core plugs extracted from the 7 core samples that were cored from Well 15/19-9 A. ***Figure 2.8*** illustrates this simple procedure.

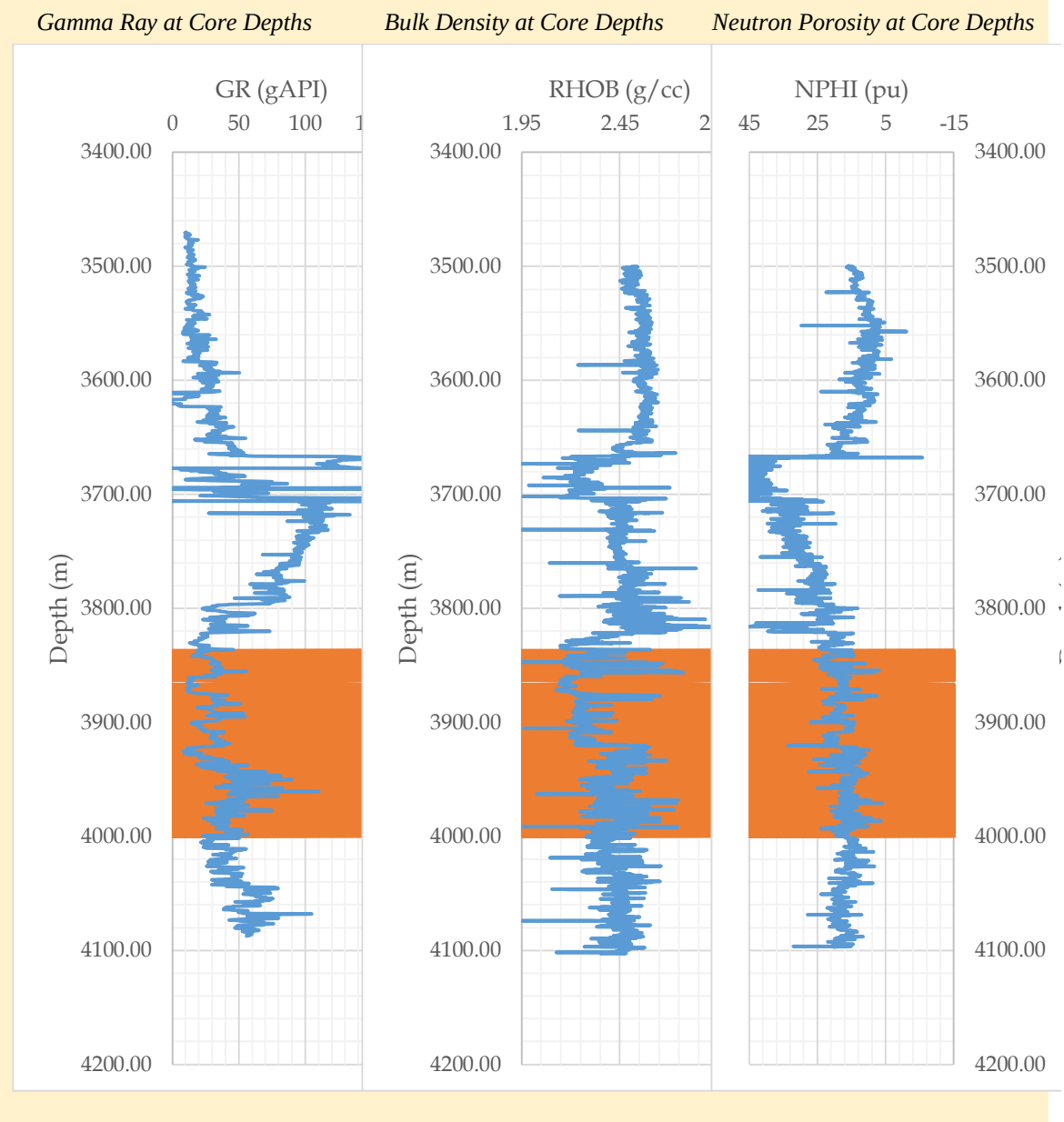


Figure 2.8 Obtaining corresponding well log values at the core depths of interest (Note: In these graphs the horizontal lines appear as an orange zone, since the lines are very close to each other).

Not all of these 728 core plugs, however, had recorded permeability values. Filtering for core plugs that underwent laboratory permeability measures reduces the number of relevant core plugs to 557. Before proceeding to training the ANN model, one would need to check for depth shift. Depth shift is a discrepancy between the log and core values in terms of the depth at which these values were measured. This difference arises due to the difference in the method of measuring the depth for both procedures. The depth for a wireline log is typically measured by use of the cable reel, whereas in coring operations, the depth would be measured via the number of drill pipes. However, since our log measurements are LWD measurements, it is generally not expected for there to be a very large disagreement. The check for depth shift can be performed by plotting log measurements with the

corresponding core measurement. ***Figure 2.9*** illustrates this check with RHOB and measured grain density from Well 15/19-9 A.

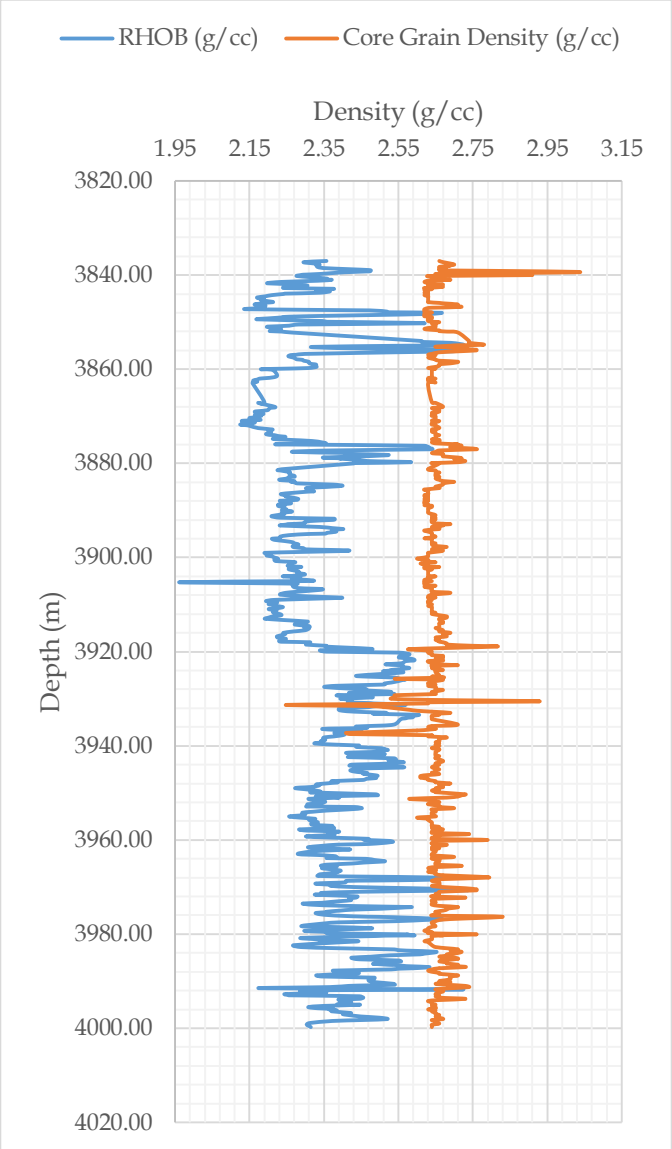


Figure 2.9 Depth shift check by plotting RHOB alongside measured grain density from Well 15/19-9 A.

The depth shift analysis indicates that the log measurements are mostly in line with the core measurements in terms of depth. Therefore, we need not concern ourselves with correcting for depth shift in this particular case. It needs to be stressed, however, that this is an important step in the QC of log data. Core values that do not correspond to the log values in terms of depth would not be able to properly train an ANN model.

A statistical analysis of the data is then performed in order to understand its general behavior. The statistical analysis of data is a process with several phases, each with its own goal. Moreover, statistical analysis provides a summary of data in the form of graphics and parameters, provides critical information to be used in estimation and conditional

simulation, and serves as an efficient vehicle for communicating information. The following are some of the statistical measures (major parameters summarized in ***Figure 2.10***) and graphic representations that were used to understand the quality of data, the quality of measurements, and the characteristics of data sample (Montgomery & Runger, 2014).

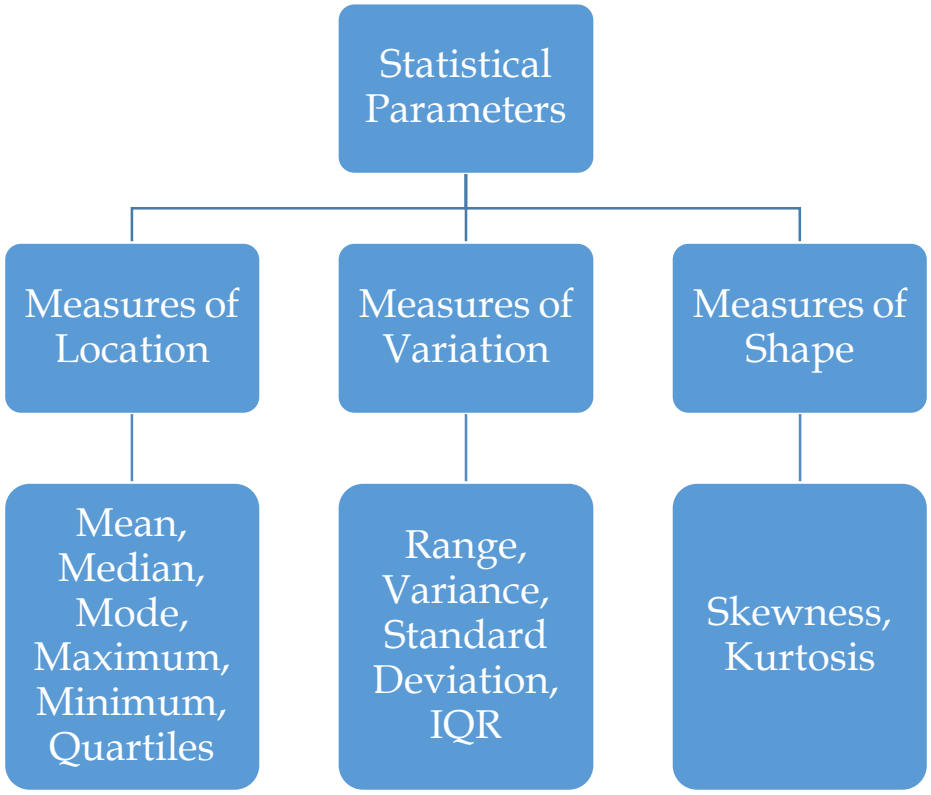


Figure 2.10 Summary of parameters to be measured in the initial statistical analysis.

Once all of the above stated procedures have been completed, the data is then ready to be input into the ANN algorithm for training. As previously mentioned, the algorithm will then tweak the weights that are multiplied to the scalar product (scalar in this particular case) and the biases that are added to that product, input this expression into the transfer function, and produce an output at the other end. We provide the output to the algorithm for 70% of the data set values (70% of a total of 557 data points) so that the algorithm can learn the trends in the log data that would result in the specified outputs. The algorithm is then tested on the remaining 30% of the data set to gauge how well it is able to predict the formation permeability. The optimization of the results is achieved by comparing three main statistical parameters that include the correlation coefficient (CC), the root mean-squared error (RMSE) and the absolute average percentage error (AAPE). ***Figure 2.11*** presents a simplification of this workflow.

The *mean squared error* (MSE) function is a performance function that calculates the average squared error between the network outputs **a** and the target outputs **t**. It is commonly used to evaluate the performance of a neural network, where the network would

be varying the weights and biases that are applied to the input to achieve a lower MSE. The goal is to minimize the average of the sum of these errors (The MathWorks, Inc., 2009).

$$mse = \frac{1}{Q}\sum_{k=1}^{Q} e(k)^2 = \frac{1}{Q}\sum_{k=1}^{Q}\big(t(k) - a(k)\big)^2 \qquad (2.1)$$

The *root mean squared error* (RMSE) function calculates the average root squared error.

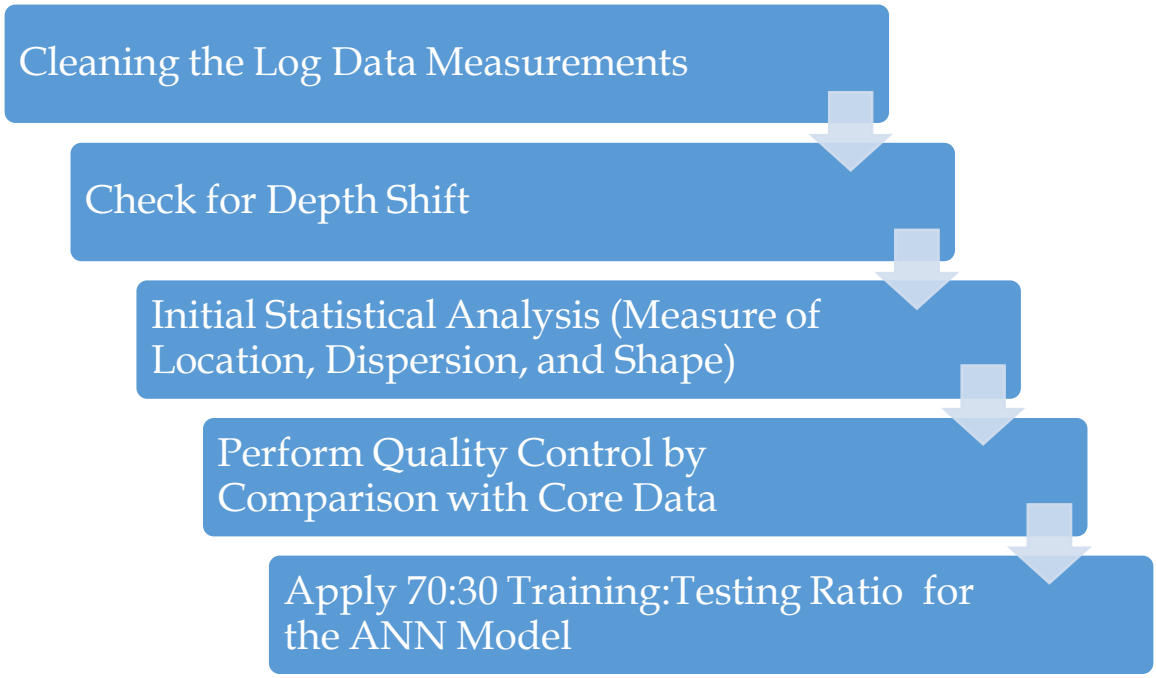


Figure 2.11 Simplified workflow for preparing the log data measurements for analysis by the ANN model.

2.4. Results and Discussion

As previously mentioned the results of the core analysis provided data on gas and liquid permeability in the horizontal and vertical directions. There are slight differences in the underlying principles by which gas and liquid permeability values are obtained, but it needs to be kept in mind that the permeability of a rock sample is an intrinsic property of the porous media itself and is independent of the nature of the fluid (at 100% saturation) that flows through the pore space (Reservoir Rock Properties Laboratory Manual, 2012). Permeability is a measure of the ease in which a fluid can pass through the pore space of a porous rock sample. In other words, it is the ability of the porous medium to transmit a fluid. The famous relationship that relates the permeability of a rock sample to the fluid that travels through it is known as Darcy's law – named after the French engineer Henry Darcy who performed a series of experiments on the flow of a fluid in a downward gradient through porous media. This relationship is given by the following expression:

$$k = \frac{Q\mu L}{A(P_1 - P_2)} \qquad (2.2)$$

In Eq. *(2.2)* k is the permeability is Darcies, μ is the viscosity of the fluid in centiPoise (cP), L is the length of the porous medium in cm, A is the cross-sectional area of flow in cm^2, P_1 is the upstream pressure in atm, and P_2 is the downstream pressure in atm.

Liquid permeability is measured in the laboratory by passing brine through the porous medium and measuring the time for the liquid to travel through it. Measurements can be taken at different flowrates to guarantee that the fluid is within the region where Darcy's law applies. Liquid permeability can then be calculated by following a modified version of Darcy's law:

$$k_l = \frac{\left(\frac{V_b}{t}\right)\mu_b L \times 10^3}{A(P_U - P_a)} \qquad (2.3)$$

In Eq. *(2.3)* k_l is the liquid permeability in mD, V_b is the brine volume in ml, t is the time in seconds for flowing 10 ml of brine through the sample, μ_b is the brine viscosity in cP, L is the length of the sample in cm, A is the cross-sectional area of the sample in cm^2, P_U is the upstream pressure in psia, and P_a is the atmospheric pressure in psia.

It should be noted that when selecting a test fluid to be passed through the porous medium, care needs to be taken to ensure the fluid will not react with the porous medium and, thereby, alter the pore structure. This will help to ensure that the measurements are accurate and reproducible after multiple tests on the same sample.

Similarly, gas permeability is measured in the laboratory by passing a gas (commonly air, helium, or nitrogen) through the pore medium and relating the flow of gas through the medium to the permeability via Darcy's law. In this case, however, the test fluid is compressible in contrast to the previous method where we would pass incompressible brine through the medium. The flowrate of the gas that passes through the medium is expressed at the mean pressure of the flow. A phenomenon that needs to be noted here is the *Klinkenberg* effect (Klinkenberg, 1941). In liquids, the molecules that are adjacent to the pore surface are stationary (no-flow conditions), whereas in gas, the molecules are in motion. This movement of the gas molecules adjacent to the surface is known as molecular slip and it is dependent on the pressure, temperature, and molecular size of the gas flowing through the medium. The gas permeability can then be found by:

$$k_g = \frac{2Q_g\mu_g L P_m \times 10^3}{A\left(P_1{}^2 - P_2{}^2\right)} \qquad (2.4)$$

In Eq. *(2.4)* k_g is the gas permeability in mD, Q_g is the gas flow rate in cm/sec, μ_g is the gas viscosity in *cP*, L is the length of the sample in cm, P_m is the mean pressure of flow in atm, A is the cross-sectional area of the sample in cm^2, P_1 and P_2 are the upstream and downstream pressures, respectively in atm.

The gas permeability can be related to the liquid permeability by use of the Klinkenberg constant and mean flow pressure in the following expression:

$$k_g = k_l\left(1 + \frac{b}{P_m}\right) \qquad (2.5)$$

where b is the Klilkenberg constant for a given gas in a given porous medium.

These slight differences in measuring principles would result in slight differences in measured permeability values. For the purposes of this study, however, these differences are so marginal that it would not matter which set of permeability values one uses to train the ANN model. Refer to ***Figure 2.12*** for a graphical representation of the liquid and gas permeability measurements obtained from Well 15/19-9 A.

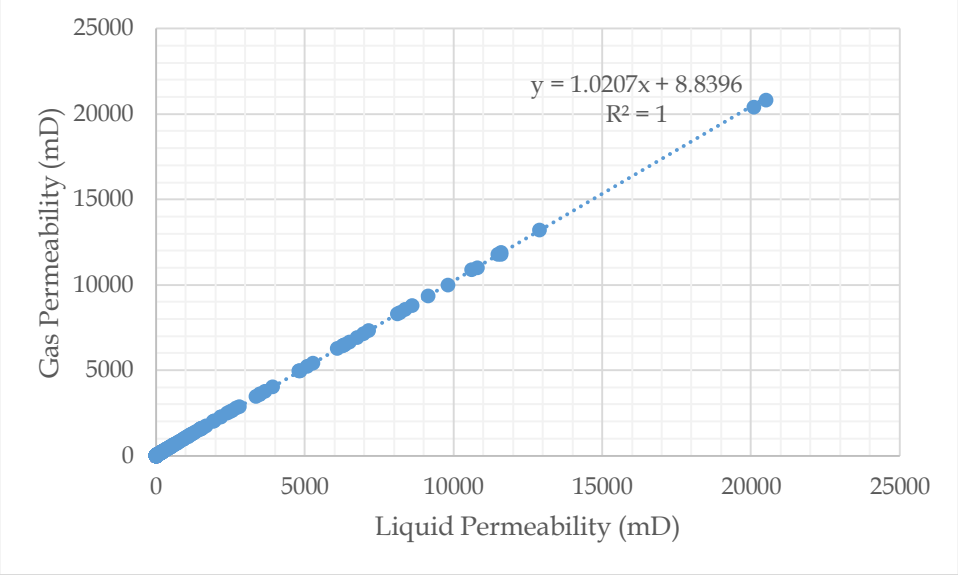


Figure 2.12 Cross-plot of gas permeability and liquid permeability values obtained from the routine core analysis performed on 557 core plugs extracted from core samples cored from Well 15/19-9 A.

As can be observed from ***Figure 2.12***, there are only slight differences between these measurements. A linear relationship can be developed which relates the gas permeability to the liquid permeability. This is given as:

$$k_g = 1.0207k_l + 8.8396 \qquad (2.6)$$

To that end, we make use of the liquid permeability values to train the ANN model. A plot of liquid permeability as a function of depth is presented in ***Figure 2.13***.

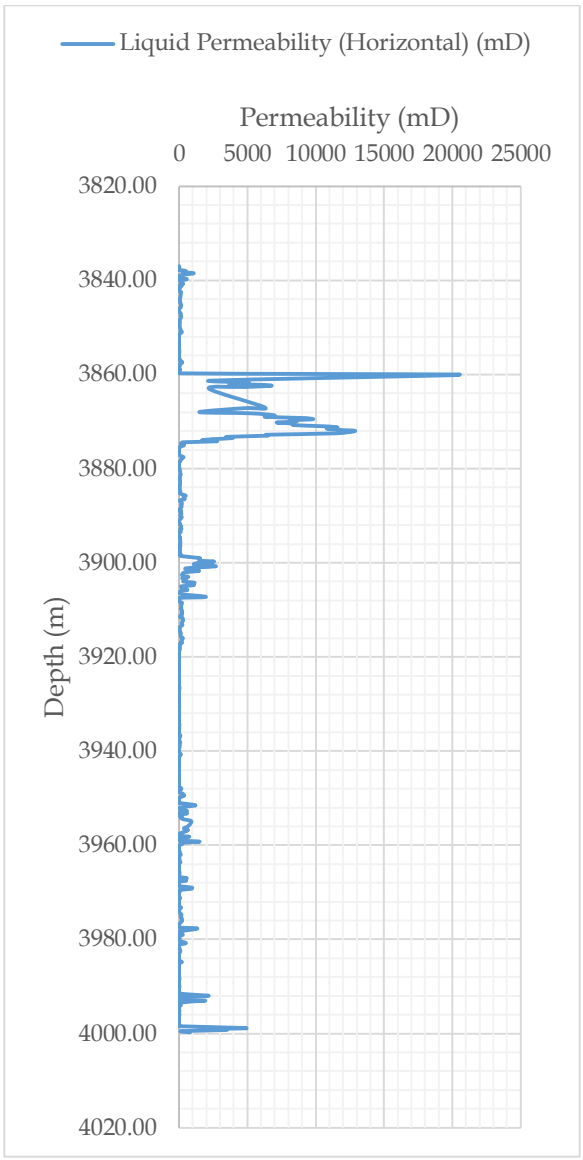
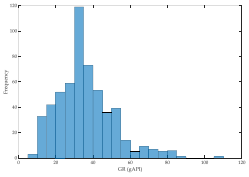
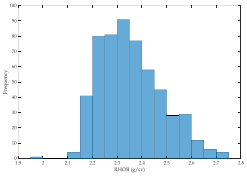
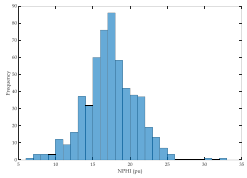


Figure 2.13 Horizontal liquid permeability data obtained from the routine core analysis performed on 557 core plugs extracted from core samples cored from Well 15/19-9 A.

Table 2.4 summarizes the statistical analysis that was performed on the three input logs (refer to ***Table 2.7*** for a simple MATLAB® code to conduct this analysis quickly). The correlation coefficient between the input data and the core permeability values were -0.3250 for gamma ray, -0.3524 for bulk density, and 0.0183 for neutron porosity. These are fairly low correlation coefficient values for all three inputs, especially for neutron porosity where the CC indicates that no correlation exists between NPHI and core permeability.

Table 2.4 Summary of statistical analysis for the three input data sets (Gamma Ray, Bulk Density, and Neutron Porosity).

Parameter	*Gamma Ray (gAPI)*	*Bulk Density (g/cc)*	*Neutron Porosity (pu)*
Minimum	8.45	1.96	6.18

Maximum	107.22	2.74	32.46
Arithmetic Mean	35.52	2.36	17.36
Geometric Mean	32.40	2.35	17.01
Harmonic Mean	29.19	2.35	16.61
Mode	8.45	1.96	6.18
Range	98.77	0.78	26.28
Mid-Range	57.83	2.35	19.32
Median	33.21	2.34	17.29
Variation	226.91	0.02	11.51
IQR	17.77	0.18	3.89
Standard Deviation	15.06	0.13	3.39
Skewness	0.93	0.55	0.04
Kurtosis	4.49	2.93	4.00
Covariance	226.91	0.02	11.51
Coefficient of Variation	0.42	0.05	0.20
Correlation Coefficient	-0.33	-0.35	0.02
Histogram Plot			

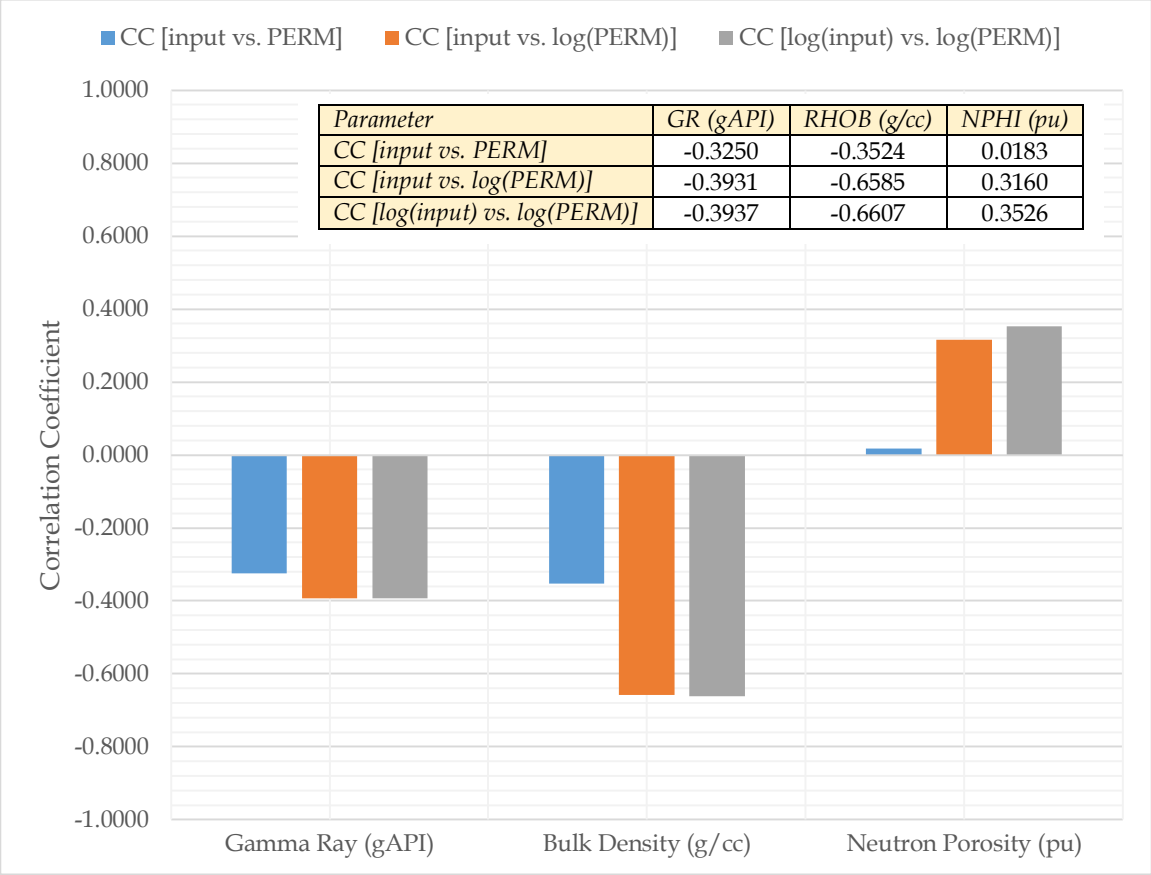
The correlation coefficients between the input parameters and the output can, however, be improved by taking the logarithm of the variables. ***Figure 2.14*** shows the result of taking the logarithm of the variables and contrasts this to the initial CC analysis without taking the logarithm. Calculating the CC for the input logs vs. the logarithm of the permeability results in CCs of -0.3931 for gamma ray, -0.6585 for bulk density, and 0.3160 for neutron porosity. These represent a percentage increase in CC of 29.8% for gamma ray, 86.9% for bulk density, and a staggering 1,626.8% increase for neutron porosity. The CC was also noted to increase further by taking the logarithm of the inputs as well as the output. These represent modest percentage increases in CC of 0.15% for gamma ray, 0.33% for bulk density, and 11.6% for neutron porosity. It is advised to perform this analysis before attempting to train the ANN model, as having inputs with a high correlation with the output would aid the neural network in converging to a more optimal solution. It should be noted, however, that a low correlation coefficient does not always mean that a relationship does not exist between the variables. Take, for instance, the simple case of the correlation coefficient of a variable x vs. a variable y which is equal to $\sin(x)$. The correlation coefficient for x vs. y for an x range of $0 \leq x \leq 100\pi$ would be equal to -0.0156 – a very low correlation coefficient, although we know there exists a direct relationship between these two. This is understandable, however, since the sine function is an oscillating one.

An attempt was made to improve the CC between the input and the output further by normalizing the input values by the following two-point slope expression:

$$Normalized\ X = 2 \times \left(\frac{X - X_{min}}{X_{max} - X_{min}}\right) - 1 \qquad (2.7)$$

where X is the input parameter (i.e. GR, RHOB, or NPHI).

This, however, did not improve the CC, in this case, since normalizing the values simply re-scales them from -1 to 1. We point it out here since it has been claimed in the literature that this can help enhance results.



Parameter	*GR (gAPI)*	*RHOB (g/cc)*	*NPHI (pu)*
CC [input vs. PERM]	-0.3250	-0.3524	0.0183
CC [input vs. log(PERM)]	-0.3931	-0.6585	0.3160
CC [log(input) vs. log(PERM)]	-0.3937	-0.6607	0.3526

Figure 2.14 Correlation coefficients between three log inputs and core permeability output.

A single layered neural network was utilized to train the ANN model. The training algorithm that was used was the Levenberg-Marquardt[3] algorithm. The stopping criteria

[3] The *Levenberg-Marquardt* algorithm was designed to approach second-order training speed without having to compute the Hessian matrix. When the performance function has the form of a sum of squares (as is typical in training feedforward networks), then the Hessian matrix can be approximated as:

for the number of iterations the algorithm will perform until it converges was set to 1,000, learning rate (a training parameter that controls the size of the weight and bias changes during learning) was set to 0.12, and the learning algorithm parameter was set to 0.6. The number of neurons used in the single layer was initially set to 20, and the transfer function used in the single hidden layer was initially selected to be the tan-sigmoid transfer function.

Neural networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the connections between elements largely determines the network function. A neural network is trained to perform a particular function by adjusting the connections between the elements (weights). Typically, neural networks are trained so that a particular output can be achieved from provided inputs. Refer to ***Figure 2.15*** for a simplified schematic of how a neural network works (The MathWorks, Inc., 2009).

$$\mathbf{H} = \mathbf{J}^T\mathbf{J}$$

and the gradient can be computed as:

$$g = \mathbf{J}^T e$$

where **J** is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights and biases, and *e* is a vector of network errors. The Jacobian matrix can be computed through a standard backpropagation technique that is much less complex than computing the Hessian matrix. The Levenberg-Marquardt algorithm uses this approximation to the Hessian matrix in the following Newton-like update:

$$x_{k+1} = x_k - [\mathbf{J}^T\mathbf{J} + \mu\mathbf{I}]^{-1}\mathbf{J}^T e$$

When the scalar μ is zero, this is just Newton's method, using the approximate Hessian matrix. When μ is large, this becomes gradient descent with a small step size. Newton's method is faster and more accurate near an error minimum, so the aim is to shift toward Newton's method as quickly as possible.

Thus, μ is decreased after each successful step (reduction in performance function) and is increased only when a tentative step would increase the performance function. In this way, the performance function is always reduced at each iteration of the algorithm (The MathWorks, Inc., 2009).

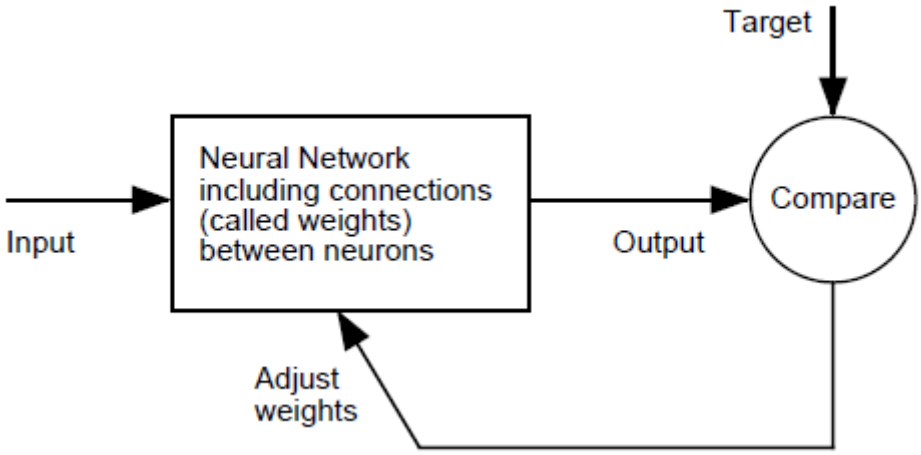


Figure 2.15 How a Neural Network works (The MathWorks, Inc., 2009).

Neural networks have been trained to perform complex functions in various fields, including pattern recognition, identification, classification, speech, vision, and control systems. Neural networks can also be trained to solve problems that are difficult for conventional computers or human beings.

Figure 2.16 shows the results of correlation coefficients obtained between the results of the computation vs. the core permeability for 100 different iterations. The iterations differ in the way the data is randomized. Out of the 557 core data points, 500 were initially randomly selected to train the model. 100 different random selections were conducted by use of a 'for loop' and the rand('seed') function available in MATLAB® to fix the random selection and ensure reproducibility.

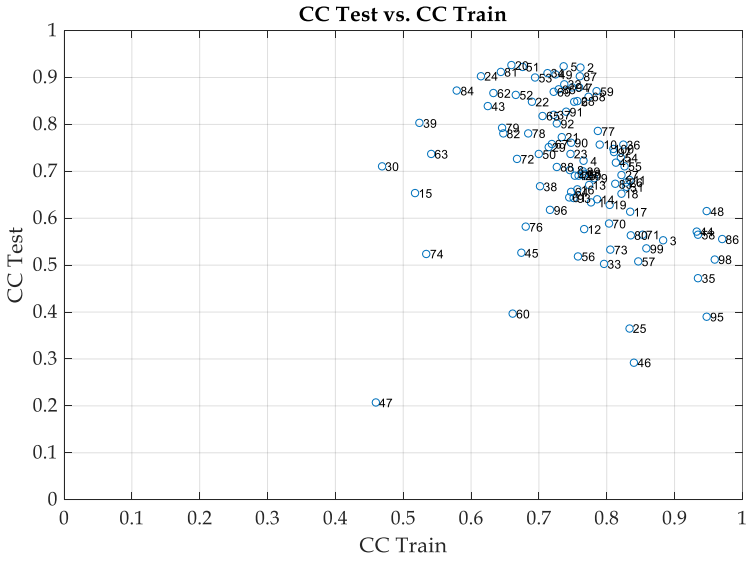


Figure 2.16 Cross-plot of the correlation coefficient between the results of the computation and the output for the case of input vs. output without taking the logarithmic value of any of the variables. The y-axis displays the

CC with the tested data points (30%), and the x-axis displays the CC with the data points used for training (70%).

As can be observed from ***Figure 2.16***, the scatter of the data is quite large with ranges of correlation coefficients that range from 0.2070 to 0.9261 for CC test, and from 0.4593 to 0.9699 for CC train. These results are undesirable since they show that the performance of the ANN model is highly sensitive on the selection of the training and testing data sets. The results may be attributed to the low CC that was displayed between the inputs and the outputs without taking the logarithmic value of any of the variables (refer to ***Figure 2.14***).

The same analysis was performed for the case of taking the logarithm of the output values before training the model. As shown in ***Figure 2.14***, taking the logarithm of the output helped significantly in improving the correlation between the inputs and the formation permeability. ***Figure 2.17*** shows the results of this analysis. The scatter has been reduced with ranges of correlation coefficients that range from 0.2855 to 0.8196 for CC test, and from 0.5877 to 0.8083 for CC train.

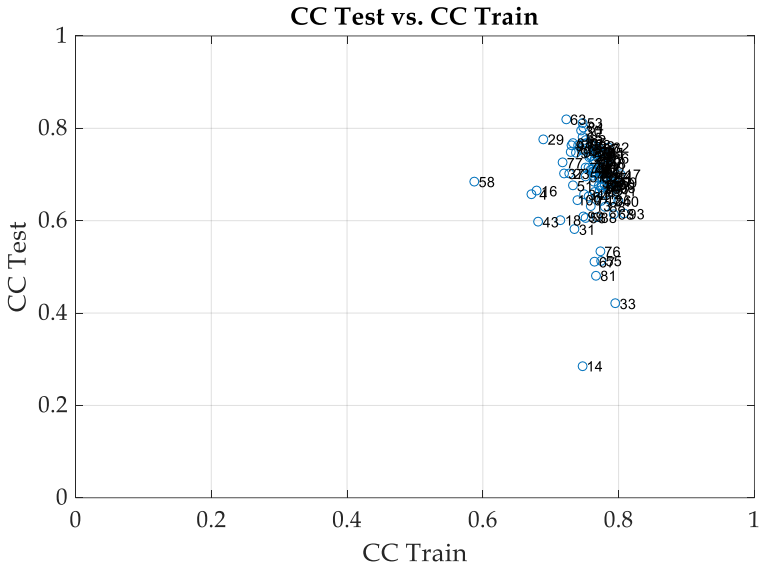


Figure 2.17 Cross-plot of the correlation coefficient between the results of the computation and the output for the case of input vs. log(output). The y-axis displays the CC with the tested data points (30%), and the x-axis displays the CC with the data points used for training (70%).

Finally, the analysis was performed for the case of taking the logarithm of both the input and output variables. ***Figure 2.18*** displays the results of this analysis. As can be observed the scatter has been significantly reduced (with the exception of point 58, which appears to be an outlier). These results show that changing the domain of the input and output variables can help in improving the correlation between the results of the computation and the output that is being predicted. The ranges of correlation coefficients range from 0.4653 to 0.8019 for CC test, and from 0.6077 to 0.8171 for CC train.

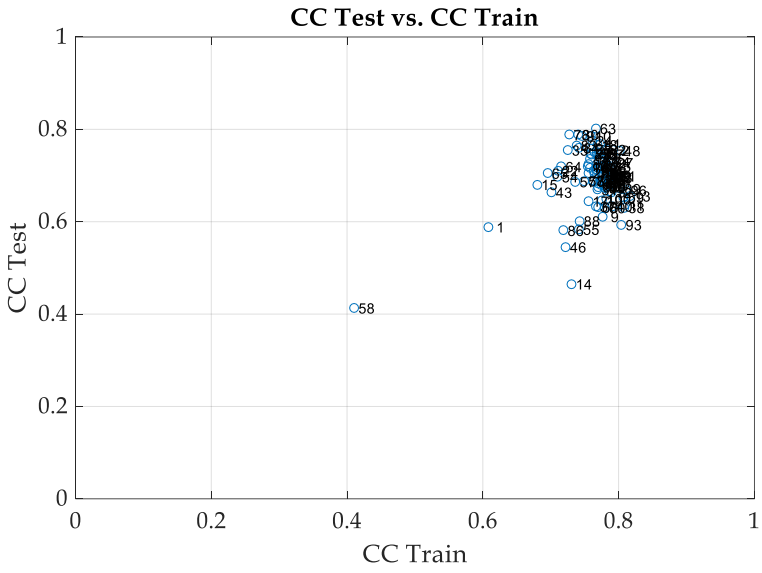


Figure 2.18 Cross-plot of the correlation coefficient between the results of the computation and the output for the case of log(input) vs. log(output). The y-axis displays the CC with the tested data points (30%), and the x-axis displays the CC with the data points used for training (70%).

Next, we conduct several different scenarios that involve varying the number of randomly selected data points from the total of 557 data points, the transfer function used to produce the output, and the number of neurons in the layer. In each case 100 iterations of different random selections were performed to determine the optimal solution. ***Table 2.5*** summarizes the results of these scenarios.

Table 2.5 Summary of the performance of different scenarios.

Case	*No. of Randomly Selected Data Points*	*No. of Neurons in Layer*	*Transfer Function*	*CC Test*	*RMSE Test*	*AAPE Test*
01	100	20	tansig	0.809	0.716	123.23
02	125	20	tansig	0.848	0.489	83.89
03	150	20	tansig	0.784	0.540	48.16
04	175	20	tansig	0.839	0.550	42.03
05	200	20	tansig	0.828	0.396	75.79
06	225	20	tansig	0.842	0.357	43.77
07	250	20	tansig	0.831	0.407	49.61
08	275	20	tansig	0.816	0.360	69.52
09	300	20	tansig	0.816	0.369	78.18
10	325	20	tansig	0.821	0.447	99.18
11	350	20	tansig	0.792	0.489	77.85
12	375	20	tansig	0.811	0.457	142.03
13	400	20	tansig	0.829	0.425	118.45
14	425	20	tansig	0.805	0.457	123.81
15	450	20	tansig	0.818	0.443	83.36
16	475	20	tansig	0.797	0.472	115.90
17	500	20	tansig	0.785	0.497	106.96
18	550	20	tansig	0.707	0.540	100.85
19	557	20	tansig	0.775	0.469	93.13
20	225	10	tansig	0.841	0.412	39.86
21	225	11	tansig	0.838	0.370	37.73

22	225	12	tansig	0.838	0.370	46.59
23	225	13	tansig	0.831	0.369	49.15
24	225	14	tansig	0.871	0.331	94.03
25	225	15	tansig	0.846	0.384	75.98
26	225	16	tansig	0.883	0.369	78.71
27	225	17	tansig	0.847	0.420	45.09
28	225	18	tansig	0.847	0.419	29.13
29	225	19	tansig	0.830	0.387	91.34
30	225	18	purelin	0.866	0.429	38.89
31	225	18	logsig	0.859	0.397	27.28

The estimated optimum number of randomly selected data points from the total of 557 was found to be 225 by holding the number of neurons in the hidden layer constant and holding the transfer function constant. The number of randomly selected data points was varied from 100 to 550 with increments of 25. A case with all of the 557 data points is also presented. The CC for the tested data was 0.842, RMSE was 0.357 and AAPE was 43.77. Next, the number of neurons was varied from 10 to 20 in increments of 1. In this case, 225 randomly selected data points were used, which is the optimum that was found from the previous analysis. The CC was 0.847, RMSE was 0.419, and AAPE was 29.13. Finally, two other transfer functions were used that included the linear transfer function (*purelin*) and the log-sigmoidal transfer function (*logsig*). The log-sigmoidal transfer function was found to perform slightly better than the other two. The CC was 0.859, RMSE was 0.397, and AAPE was 27.28. It should be noted that varying the stated parameters in these scenarios may help to slightly tweak the final results, however, the majority of the differences in prediction results actually comes from the initial setting up of the problem at hand. All in all, a total of 31 cases were performed, where the best results out of the 100 iterations would be recorded. ***Figure 2.19*** illustrates the results of the regression analysis and the performance of the algorithm for this best-case scenario (*Case 31*). Moreover, ***Figure 2.20*** shows the logarithm of the predicted and original permeability as a function of depth for this scenario.

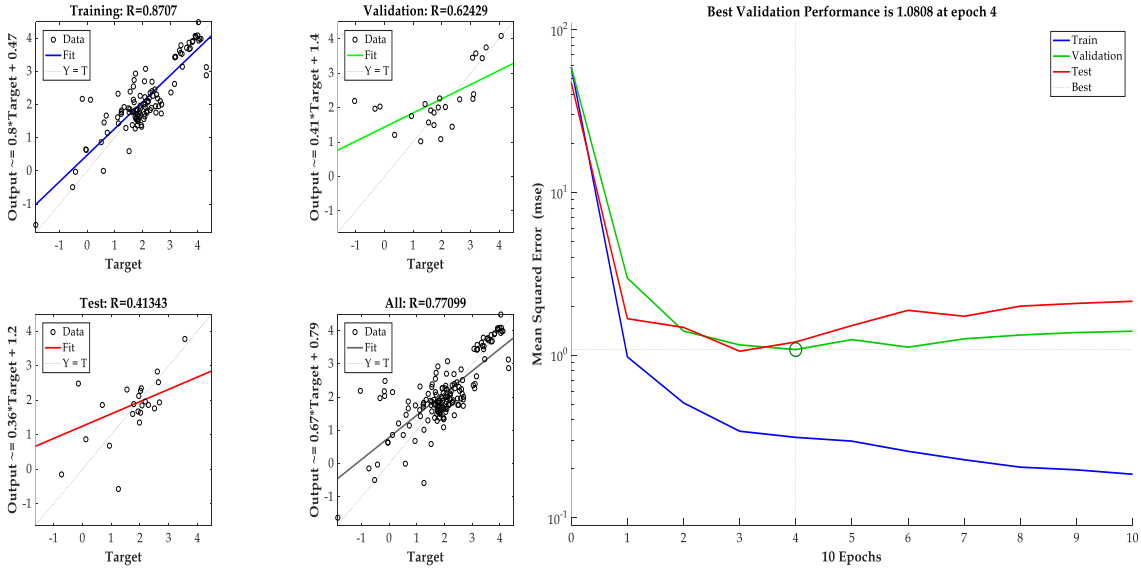


Figure 2.19 Left: Regression analysis for Case 31 (30% Training, 15% Validation, 15% Testing); Right: Performance of the algorithm for Case 31. The performance was gauged by attempting to minimize the mean squared error (MSE).

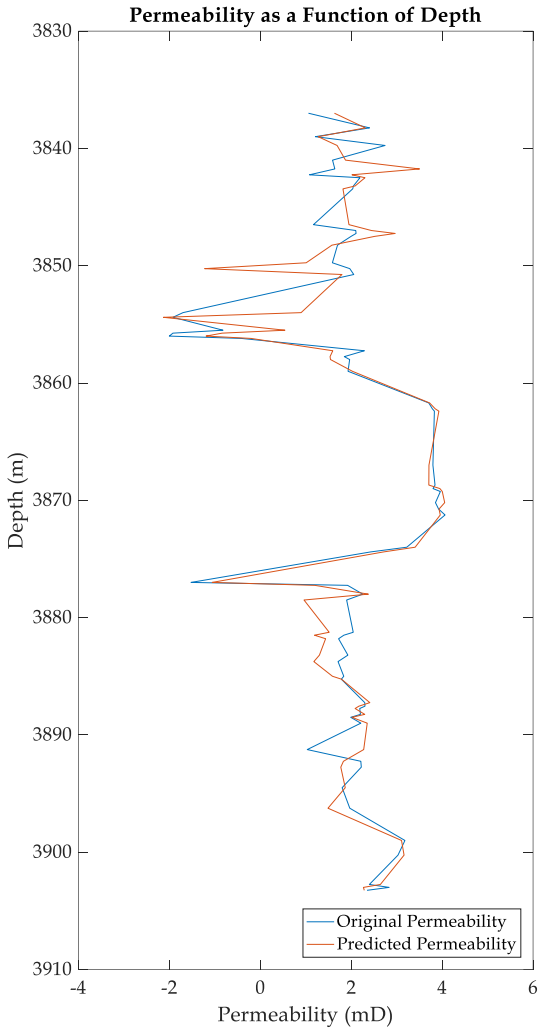


Figure 2.20 Predicted and original permeability as a function of depth for Case 31.

The weights and biases used in the developed ANN model may then be extracted from the network (refer to ***Table 2.8*** for the MATLAB® code that can be used to do this). ***Table 2.6*** tabulates the extracted weights for the input and hidden layer, and the extracted biases for the input and the output layer. These values may then be used to calculate the logarithm of the predicted permeability (k) by the use of the following mathematical expression:

$$\log(k) = \left[\sum_{j=1}^{N} w_{2i}\, logsig\left(\sum_{j=1}^{J} w_{1i,j}x_j + b_{1i}\right)\right] + b_2 \qquad (2.8)$$

where i = 1 represents the logarithm of the gamma ray log values, i = 2 represents the logarithm of the bulk density log values, i = 3 represents the

logarithm of the neutron porosity log values, *N* is the number of neurons (18 in this case), and *J* is the number of input variables (3 in this case).

Table 2.6 Extracted weights and biases from the developed ANN model.

Input Layer Weight Matrix			*Input Layer Bias Vector*	*Hidden Layer Weight Vector*	*Output Layer Bias Vector*
w_{11}	w_{12}	w_{13}	b_1	w_2	b_2
4.7563	2.0720	-4.7042	-7.0791	-1.0529	-1.0586
1.0714	1.5857	6.9763	-6.4504	-0.2996	
-4.5380	5.7318	-2.7601	4.9276	0.1089	
-4.0998	-1.3061	5.7180	4.9286	0.4484	
2.3120	0.7708	-8.2534	-1.6330	2.9098	
-1.0398	3.4567	-6.4098	3.1673	0.0411	
5.2402	4.9907	-3.4092	-2.5322	0.2315	
2.4164	3.0799	-5.5372	-0.8083	-3.5548	
-3.3598	3.6852	-6.1119	-0.2335	-0.3795	
-6.1331	-2.8086	4.5172	1.2980	-0.4742	
3.6356	-8.9837	-2.0296	0.3280	-0.4747	
-5.2739	4.9948	-1.2020	-2.4155	0.2957	
-8.4728	-0.9722	2.1388	-4.3755	-0.5340	
-5.2969	-1.5623	-5.4275	-2.5612	0.2898	
0.0984	-0.6479	7.3457	-5.5522	-0.4225	
-2.4484	-6.6114	1.4537	-5.5068	1.6601	
4.9370	-1.0529	5.0538	6.5238	0.9963	
1.7470	8.2357	-1.4964	7.2396	0.6986	

2.5. Conclusion

To conclude, the work in this paper attempted to predict formation permeability obtained from core plugs in an exploratory well drilled in the Volve field located in the Norwegian continental shelf via the use of a single layered ANN model. The input data that was fed into the network were three common logs that include the gamma ray (GR) log, the bulk density (RHOB) log, and the neutron porosity (NPHI) log. It has been shown that taking the logarithm of both the input data and the output data, in this case, would improve the results of the prediction. An attempt at normalizing the input data to range between -1 and 1 using a two-point slope did not improve results of the prediction. The optimum single layered ANN model was determined to be one with 18 neurons in the hidden layer, and a log-sigmoidal transfer function. The optimum number of randomly selected data points from the total of 557 data points was found to be 225. An important observation to be made is that the initial setting up of the problem is more significant than varying parameters in the network, such as the number of neurons in the hidden layer, and the transfer function used in the hidden layer, in achieving good prediction results.

A reasonable cut-off value for permeability for CO_2 sequestration is 10 mD. A region or interval with a permeability greater than 10 mD would allow for CO_2 to easily travel to the intended area of accumulation. This cut-off value, however, will ultimately be decided by the party intending to sequester CO_2.

This paper presents a systematic methodology to predict core formation permeability from well log data using ANN for the purposes of characterizing a potential CO_2 sequestration site. It is hoped that this work may pave the path for practitioners to adopt a comprehensive screening procedure in locating a suitable CO_2 sequestration site by use of artificial intelligence tools (such as ANN) and massive amounts of data already available from the Upstream Oil & Gas Industry.

2.6. Future Work

As previously mentioned, the aspects of locating a CO_2 sequestration site may be simplified to the following main aspects:

- **Locating your Reservoir**: (i) Good Quality Lithologies; (ii) Favorable Permeability; (iii) Reservoir Extension; (iv) Net-to-Gross Ratio.
- **Locating your Seal**: (i) Typical Seal Type Lithologies; (ii) Low Permeability Layers; (iii) Seal Extension over Reservoir; (iv) Sufficient Thickness.
- **Locating your Trap**: (i) Presence of Structures; (ii) Presence of Stratigraphic Traps; (iii) Presence of Four-Way Dip Closures; (iv) Presence of Three-Way Dip Closures.

In this paper, the discussion was focused on the determination of favorable permeability for the sake of classifying a reservoir. The authors now plan to continue with the next step, which would be the prediction of lithologies that are generally associated with good reservoirs by the use of data available from core descriptions and descriptions of cuttings available from mud logs.

2.7. Acknowledgements

The authors would like to thank King Fahd University of Petroleum & Minerals for permission to present this paper. Moreover, the first author would like to express gratitude towards his friends and colleagues in the *Artificial Intelligence in Petroleum Engineering* course, which was offered in the spring semester of 2018, for the tips and suggestions they provided as regards to the research methodology adopted in this paper.

2.8. Appendix

Table 2.7 MATLAB code used to conduct statistical analysis on input data.

```
%Statistical Analysis for Well 15_19-19 A
GR = xlsread('Well15_9_19A_Analysis.xlsx','Training-Testing Dataset','W3:W559');
RHOB = xlsread('Well15_9_19A_Analysis.xlsx','Training-Testing Dataset','X3:X559');
NPHI = xlsread('Well15_9_19A_Analysis.xlsx','Training-Testing Dataset','Y3:Y559');
PERM = xlsread('Well15_9_19A_Analysis.xlsx','Training-Testing Dataset','Z3:Z559');

%Statistical Analysis for GR
GR_Minimum = min(GR);
GR_Maximum = max(GR);
GR_ArithmeticMean = mean(GR);
GR_GeometricMean = geomean(GR);
GR_HarmonicMean = harmmean(GR);
GR_Mode = mode(GR);
```

```
GR_Range = range(GR);
GR_MidRange = (GR_Maximum + GR_Minimum)/2;
GR_Median = median(GR);
GR_Variation = var(GR);
GR_IQR = iqr(GR);
GR_StdDev = std(GR);
GR_Skewness = skewness(GR);
GR_Kurtosis = kurtosis(GR);
GR_FrequencyHistogram = histogram(GR);
GR_Covariance = cov(GR);
GR_CoeffVariation = GR_StdDev/GR_ArithmeticMean;
GR_CorrCoeff = corr2(GR,PERM);
GR_CorrCoeff2 = corr2(GR,log10(PERM));
GR_CorrCoeff3 = corr2(log10(GR),log10(PERM));

%Statistical Analysis for RHOB
RHOB_Minimum = min(RHOB);
RHOB_Maximum = max(RHOB);
RHOB_ArithmeticMean = mean(RHOB);
RHOB_GeometricMean = geomean(RHOB);
RHOB_HarmonicMean = harmmean(RHOB);
RHOB_Mode = mode(RHOB);
RHOB_Range = range(RHOB);
RHOB_MidRange = (RHOB_Maximum + RHOB_Minimum)/2;
RHOB_Median = median(RHOB);
RHOB_Variation = var(RHOB);
RHOB_IQR = iqr(RHOB);
RHOB_StdDev = std(RHOB);
RHOB_Skewness = skewness(RHOB);
RHOB_Kurtosis = kurtosis(RHOB);
RHOB_FrequencyHistogram = histogram(RHOB);
RHOB_Covariance = cov(RHOB);
RHOB_CoeffVariation = RHOB_StdDev/RHOB_ArithmeticMean;
RHOB_CorrCoeff = corr2(RHOB,PERM);
RHOB_CorrCoeff2 = corr2(RHOB,log10(PERM));
RHOB_CorrCoeff3 = corr2(log10(RHOB),log10(PERM));

%Statistical Analysis for NPHI
NPHI_Minimum = min(NPHI);
NPHI_Maximum = max(NPHI);
NPHI_ArithmeticMean = mean(NPHI);
NPHI_GeometricMean = geomean(NPHI);
NPHI_HarmonicMean = harmmean(NPHI);
NPHI_Mode = mode(NPHI);
NPHI_Range = range(NPHI);
NPHI_MidRange = (NPHI_Maximum + NPHI_Minimum)/2;
NPHI_Median = median(NPHI);
NPHI_Variation = var(NPHI);
NPHI_IQR = iqr(NPHI);
NPHI_StdDev = std(NPHI);
NPHI_Skewness = skewness(NPHI);
NPHI_Kurtosis = kurtosis(NPHI);
NPHI_FrequencyHistogram = histogram(NPHI);
NPHI_Covariance = cov(NPHI);
NPHI_CoeffVariation = NPHI_StdDev/NPHI_ArithmeticMean;
NPHI_CorrCoeff = corr2(NPHI,PERM);
```

```
NPHI_CorrCoeff2 = corr2(NPHI,log10(PERM));
NPHI_CorrCoeff3 = corr2(log10(NPHI),log10(PERM));
```

Table 2.8 MATLAB code used to extract weights and biases from developed ANN model.

```
save net;
weights_biases = getwb(net);
[b1,IW,LW] = separatewb(net,weights_biases);
```

2.9. References

Abdulraheem, A., Sabakhi, E., Ahmed, M., Vantala, A., Raharja, I., & Korvin, G. (2007). Estimation of Permeability from Wireline Logs in a Middle Eastern Carbonate Reservoir using Fuzzy Logic. *15th SPE Middle East Oil & Gas Show and Conference* (pp. 1-11). Bahrain International Exhibition Centre, Kingdom of Bahrain: Society of Petroleum Engineers.

Anifowose, F. A., Abdulraheem, A., Al-Shuhail, A., & Schmitt, D. P. (2013). Improved Permeability Prediction from Seismic and Log Data using Artificial Intelligence Techniques. *SPE Middle East Oil and Gas Show and Conference* (pp. 1-7). Manama, Bahrain: Society of Petroleum Engineers.

Bachu, S. (2000). Sequestration of CO2 in Geological Media: Criteria and Approach for Site Selection in Response to Climate Change. *Energy Conversion & Management*, 953-970.

Bachu, S. (2002). Sequestration of CO2 in Geological Media in Response to Climate Change: Road Map for Site Selection using the Transform of the Geological Space into the CO2 Phase Space. *Energy Conversion & Management*, 87-102.

Bagheripour, P. (2014). Committee neural network model for rock permeability prediction. *Journal of Applied Geophysics*, 142-148.

Baines, S. J., & Worden, R. H. (2004). Geological storage of carbon dioxide. *Geological Society, London, Special Publications*, 1-6.

Benson, S. M., & Franklin M. Orr, J. (2008). Carbon Dioxide Capture and Storage. *MRS Bulletin*, 303-305. doi:10.1557/mrs2008.63

Chadwick, R., Zweigel, P., Gregersen, U., Kirby, G., Holloway, S., & Johannessen, P. (2004). Geological Reservoir Characterization of a CO2 Storage Site: The Utsira Sand, Sleipner, Northern North Sea. *Energy*, 1371-1381.

Cuddy, S. J. (1997). The Applications of the Mathematics of Fuzzy Logic

to Petrophysics. *SPWLA 38th Annual Logging Symposium* (pp. 1-14). Houston, Texas: Society of Petrophysicists and Well Log Analysts.

Cuddy, S. J. (2000). Litho-Facies and Permeability Prediction from Electrical Logs using Fuzzy Logic. *SPE Reservoir Eval. & Eng., 3*(4), 319-324.

Doke, J. (2016, September 1). *MathWorks®*. Retrieved from File Exchange: https://www.mathworks.com/matlabcentral/fileexchange/7173-grabit

Elkatatny, S., Mahmoud, M., Tariq, Z., & Abdulraheem, A. (2018). New insights into the prediction of heterogeneous carbonate reservoir permeability from well logs using artificial intelligence network. *Neural Computing and Application*, 2673-2683.

Equinor. (2018, May 21). *Volve data village*. Retrieved July 26, 2018, from Equinor: https://data-equinor-com.azurewebsites.net/dataset/volve

Fang, Y., Baojun, B., Dazhen, T., Dunn-Norman, S., & Wronkiewicz, D. (2010). Characteristics of CO2 Sequestration in Saline Aquifers. *Petroleum Science*, 83-92.

Intergovernmental Panel on Climate Change. (2005). *Special Report on Carbon Dioxide Capture and Storage.* Cambridge, UK: Cambridge University Press.

Iturrarán-Viveros, U., & Parra, J. O. (2014). Artificial Neural Networks applied to estimate permeability, porosity and intrinsic attenuation using seismic attributes and well-log data. *Journal of Applied Geophysics*, 45-54.

Jamshidian, M., Hadian, M., Zadeh, M. M., Kazempoor, Z., Bazargan, P., & Salehi, H. (2015). Prediction of free flowing porosity and permeability based on conventional well logging data using artificial neural networks optimized by Imperialist competitive algorithm - A case study in the South Pars gas field. *Journal of Natural Gas Science and Engineering*, 89-98.

Klinkenberg, L. J. (1941). The Permeability of Porous Media to Liquids and Gases. *Production Practice*, 200-213.

Kovscek, A. R. (2002). Screening Criteria for CO2 Storage in Oil Reservoirs. *Petroleum Science and Technology*, 841-866.

Lim, J.-S., & Kim, J. (2004). Reservoir Porosity and Permeability Estimation from Well Logs using Fuzzy Logic and Neural Networks. *SPE Asia Pacific Oil and Gas Conference and Exhibition* (pp. 1-9). Perth, Australia: Society of Petroleum Engineers.

Montgomery, D. C., & Runger, G. C. (2014). *Applied Statistics and Probability for Engineers* (6th

ed.). Singapore: John Wiley & Sons Singapore Pte. Ltd.

Nashawi, I., & Malallah, A. (2010). Permeability Prediction from Wireline Logs using Fuzzy Logic and Discriminant Analysis. *SPE Asia Pacific Oil & Gas Conference and Exhibition* (pp. 1-12). Brisbane, Queensland, Australia: Society of Petroleum Engineers.

Norwegian Petroleum Directorate. (2018). *Factpages*. Retrieved from Norwegian Petroleum Directorate: http://factpages.npd.no/factpages/

Olatunji, S., Selamat, A., & Abdulazeez, A. (2015). Harnessing the Power of Type-2 Fuzzy Logic System in the Prediction of Reservoir Properties. *SPE Saudi Arabia Section Technical Symposium and Exhibition* (pp. 1-24). Al-Khobar, Saudi Arabia: Society of Petroleum Engineers.

Rafik, B., & Kamel, B. (2017). Prediction of permeability and porosity from well log data using the nonparametric regression with multivariate analysis and neural network, Hassi R'Mel Field, Algeria. *Egyptian Journal of Petroleum*, 763-778.

Reservoir Rock Properties Laboratory Manual. (2012). Dhahran: King Fahd University of Petroleum & Minerals.

Saljooghi, B. S., & Hezarkhani, A. (2015). A new approach to improve permeability prediction of petroleum reservoirs using neural network adaptive wavelet (wavenet). *Journal of Petroleum Science and Engineering*, 851-861.

Taghavi, A. (2005). Improved Permeability Estimation through use of Fuzzy Logic in a Carbonate Reservoir from Southwest Iran. *14th SPE Middle East Oil & Gas Show and Conference* (pp. 1-9). Bahrain International Exhibition Centre, Bahrain: Society of Petroleum Engineers.

The MathWorks, Inc. (2009). *Neural Network Toolbox™ 6 User's Guide*. The MathWorks, Inc.

3. CHAPTER 3 - LITHOLOGY CLASSIFICATION VIA ARTIFICIAL NEURAL NETWORKS (ANN)

3.1. Abstract

The lithology of a formation plays an important part in the rock properties that would be present in that particular formation. These lithology appear in nature as a result of the several depositional environments that the sediments were originally deposited in. It would vary greatly depending on factors such as the climate, distance the sediments travelled, whether the environment was marine or non-marine, etc. Depending on the way the sediments were deposited, properties such as permeability and porosity would be affected. These properties are crucial in the evaluation of a potential CO_2 sequestration site. The use of ANN to determine/predict favorable lithology such as sandstones and carbonates may indeed be an added advantage to governments seeking to store CO_2 underground to curb emissions.

In this paper, we focus on the use of well log data obtained from a well in the Volve field present in the Norwegian Continental Shelf to predict the lithology of the respective intervals. 557 sample points from three well logs (GR, RHOB, and NPHI) were used to classify the target lithology. Our main objective is to classify the inputs into a set of target categories. Two main tools are utilized to analyze the result of the neural network: (i) *Confusion Matrix* and (ii) *Receiver Operating Characteristic* curve. The neural network was constructed by use of a two-layer feed-forward network and was trained via the scaled conjugate back-propagation algorithm. The training dataset consisted of 70% of the data, whereas the validation and testing datasets were both 15% of the data. To this end, the MATLAB® ANN Classification GUI was utilized.

Results show that an overall classification percentage of 61.8% was achieved (by looking at the confusion matrix). Moreover, a higher true positive rate was obtained by the Dolomite class (by analyzing the ROC curve).

3.2. Introduction

In this section, we attempt to employ the Neural Network Pattern Recognition application to use a number of inputs of GR, RHOB, and NPHI to identify the lithology that correspond to them. In pattern recognition problems, the main aim is to be able to classify the inputs into a set of target categories. We adopt an approach that utilizes a tool known as the *confusion matrix* as well as the *receiver operating characteristic* curve to study the performance of the training algorithm.

In its simplest format, a *confusion matrix* is a two-dimensional matrix that summarizes the performance of a classifier. It does this by plotting the actual vs. assigned classes of a system being analyzed. For instance if a system consists of 15 cats, 10 dogs and 22 birds, a simple confusion matrix may be sketched as shown in ***Table 3.1*** (Sammut & Webb, 2017).

Table 3.1 Simple Confusion Matrix.

Actual Class	Assigned Class			
		Cats	*Dogs*	*Birds*
	Cats	12	2	1
	Dogs	2	8	0
	Birds	4	3	15

In this case,

i. the classifier correctly classified 12 out of 15 samples to be cats, however 2 and 1 were misclassified as dogs and birds, respectively;
ii. correctly classified 8 out of 10 samples to be dogs, however 2 samples were misclassified as cats;
iii. correctly classified 15 out of 22 samples to be birds, however 4 and 3 samples were misclassified as cats and dogs, respectively.

Another form of the confusion matrix that is often utilized, especially in the field of machine learning and statistical classification, is one in which the samples may either be labelled as *true positives (TP)*, *true negatives (TN)*, *false positives (FP)* or *false negatives (FN)*. Take for instance a system having 10 Red balls and 10 Blue balls,

i. if the classifier correctly classifies 8 balls as being Red, then we have 8 true positives (TP) and 2 false negatives (FN).
ii. if the classifier correctly classifies 7 balls as being Blue, then we have 7 true negatives (TN) and 3 false positives (FP).

This method is utilized when there only exists two classes in the system (i.e. Red and Blue). Refer to ***Table 3.2*** (Sammut & Webb, 2017).

Table 3.2 Specialized Confusion Matrix.

Actual Class	Assigned Class		
		Positive	*Negative*
	Positive	TP	FN
	Negative	FP	TN

Researchers have been attempting to classify formation lithology for several decades via several means of analysis on data such as core and log data. The earlier methods utilized statistical means to attempt to derive patterns in log data that may be characteristic of a particular lithology type. These methods have now been mostly replaced with artificial intelligence techniques that utilize tools such as fuzzy logic, artificial neural networks, support vector machines, etc.

As regards to statistical methods, for instance, Busch, Fortney and Berry (1987) attempted to predict lithology by the use of a statistical methodology applied to wireline log measurements that were calibrated to a core standard. This methodology was tested and applied to the Shublik formation, however the authors claim that the methodology may be

applied to any formation provided that a core standard is available to calibrate the log measurements. The authors narrowed down on the most successful model by use of statistical discriminant analysis. The final model is capable of predicting lithology at a 75% accuracy rate. Tang and White (2008) made use of three multivariate statistical methods that were beta-Bayesian, multinomial logistic regression, and discriminant analysis to predict the sedimentary facies of a western African sandstone reservoir. The method uses empirical beta distributions to model the variation and spread of petrophysical properties that are characteristic of a particular sedimentary facies. The authors were able to achieve a final prediction accuracy of 82 to 90%.

Statistical methods were replaced with artificial intelligence techniques due to the increasing demand to provide better and more efficient generalizations of models aiming to classify lithology from relevant data points. The remainder of this section will concern work done on the use of artificial intelligence techniques to classify lithology.

Al-Anazi and Gates (2010) studied the performance of a support vector machine based lithology classification methodology and contrasted its performance with that of conventional statistical methods such as discriminant analysis and probabilistic neural networks. By applying the technique to well log data points belonging to a heterogeneous sandstone reservoir, and utilizing detailed core descriptions from a training well, the authors found that the SVM-based methodology is superior in comparison with the other two methods. The proposed reason for the superior performance is the fact that the SVM methodology employs the use of statistical learning theory which achieves better generalization by minimizing the testing error. Benaouda *et al*. (1999) utilize a neural-net based technique to classify lithology using available geophysical downhole data from wells that have partial core recovery. The technique allows for better lithology generalization for areas of the wellbore where core samples are not available. The performance is compared to that of discriminant analysis and in qualitative correlation with non-depth-matched core, and was found to perform satisfactorily. Analysis was conducted on a borehole that was drilled as part of an Ocean Drilling Program. The formation penetrated was a mixture of volcaniclastic sandstones, conglomorates and sandstones. Walls *et al*. (1999) developed a method employing a combination of rock physics modeling, seismic attribute generation and pattern recognition via a neural network analysis to classify lithology by the use of core, log, and seismic data. Kapur *et al*. (1998) used a back propagation neural network to classify facies in uncored wells by use of relationships developed from cored wells. The network was trained using log data such as gamma ray, density, neutron and resistivity logs. Authors were able to achieve a 75% to 93% prediction rate. Saggaf and Nebrija (2000) made use of neural networks that perform vector quantization of input via competitive learning to identify facies in horizontal wells present in Saudi Arabia. Horizontal wells are almost never cored, so the method proves very useful in using data obtained from nearby cored wells to predict the lithology in the uncored well. Bhatt and Helle (2002) utilized a method that relies on back propagation neural networks in ensembles and modular systems to aid in the identification of facies. The model was developed using synthetic data and it was used to in the Ness formation of the North Sea.

More recently, Ghosh, Chatterjee and Shanker (2016) attempted to predict coal facies from five exploratory wells by use of linear regression analysis and multilayer feed forward neural network models. Data used comprised of natural gamma ray, high resolution density and single point resistivity to determine the ash % and moisture % in coal layers. Authors were able to predict the coal facies (i.e. coal, shaly coal and shale) in three testing wells using the other two wells for training and validation. Ojha and Maiti (2016) utilized a combination of a Bayesian neural network and Markov Chain Monte Carlo/hybrid Monte Carlo learning paradigm. The method was used to classify sediment/lithology boundaries by use of data such as gamma ray, sonic P-wave, porosity, density porosity and electrical resistivity in a borehole drilled as part of the Integrated Ocean Drilling Program (IODP). Results show that the BNN based method is robust in its classification of lithology boundaries. Imamverdiyez and Sukhostat (2019) proposed a new convolutional neural network (CNN) that was trained on a variety of optimization algorithms to classify geologic facies. The model was compared to 4 other models and was shown to outperform them. Inputs used to train the model include gamma ray, photoelectric effect, resistivity and average neutron porosity. Deng *et al*. (2017) propose the use of an SVM to classify crystalline rocks. Moreover, the authors introduce a synthetic minority over-sampling techniques (SMOTE) and Borderline-SMOTE to deal with imbalanced data that has been shown to affect the classification results. The performance of the model was compared to that of a BPNN. Son *et al*. (2016) also applied ANN to determine porosity as well as classify lithological facies by use of core and log data. Results show that error of estimated porosity ranges from -0.3 to 0.3 and accuracy of lithology classification is 70%. Silva *et al*. (2015) use ANN to classify lithology of carbonate-siliciclastic rocks using BPNN. Data used include elastic, mineralogical, and textural information from well located in the southwest of France. Mahmoodi, Smith and Tinkham (2016) utilized neural networks to predict rock types by use of density, gamma ray, and magnetic susceptibility measurements. Results show that model was able to achieve a predictive accuracy of 83%. Moreover, model is better able to classify homogeneous rock types in comparison to heterogeneous rock types.

More methods of classifying lithology from well logs are also available from Ma (2019).

3.3. Methodology

The same filtered data set of 557 samples from the previous chapter were used in approaching this problem. Recall that all of these 557 samples were selected based on available core permeability data. The interval that we are studying, therefore, is the cored interval of Well 15/9-19 A (3837 – 4000 m). A segment of the formation evaluation log where these lithology were obtained is shown in ***Figure 3.2***. Throughout this interval, five main lithology classes were observed. Each of these classes were given a lithology code in order to distinguish them from each other. ***Table 3.3*** summarizes this.

Table 3.3 Lithology codes for respective lithology classes.

Lithology	*Lithology Code*
SST (SANDSTONE)	10000
DOL (DOLOMITE)	01000

CLST (CLASTIC)	00100
COAL (COAL)	00010
SLTST (SILTSTONE)	00001

The neural network was constructed by use of a two-layer feed-forward (i.e. a layered network in which each layer only receives inputs from previous layers) network (see ***Figure 3.1***), with sigmoid hidden and softmax output neurons (patternnet). This network can classify vectors arbitrarily well, given enough neurons in its hidden layer. The network will be trained with scaled conjugate gradient backpropagation (trainscg).

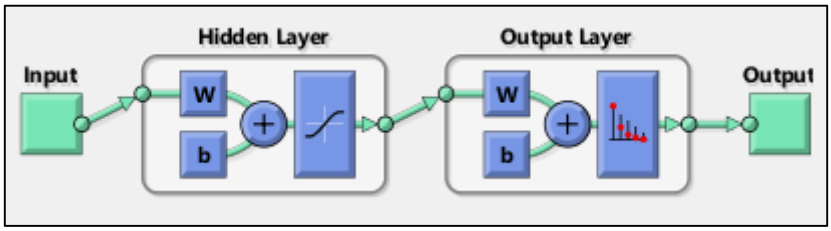


Figure 3.1 Two-layer feed-forward network, with sigmoid hidden and softmax output neurons.

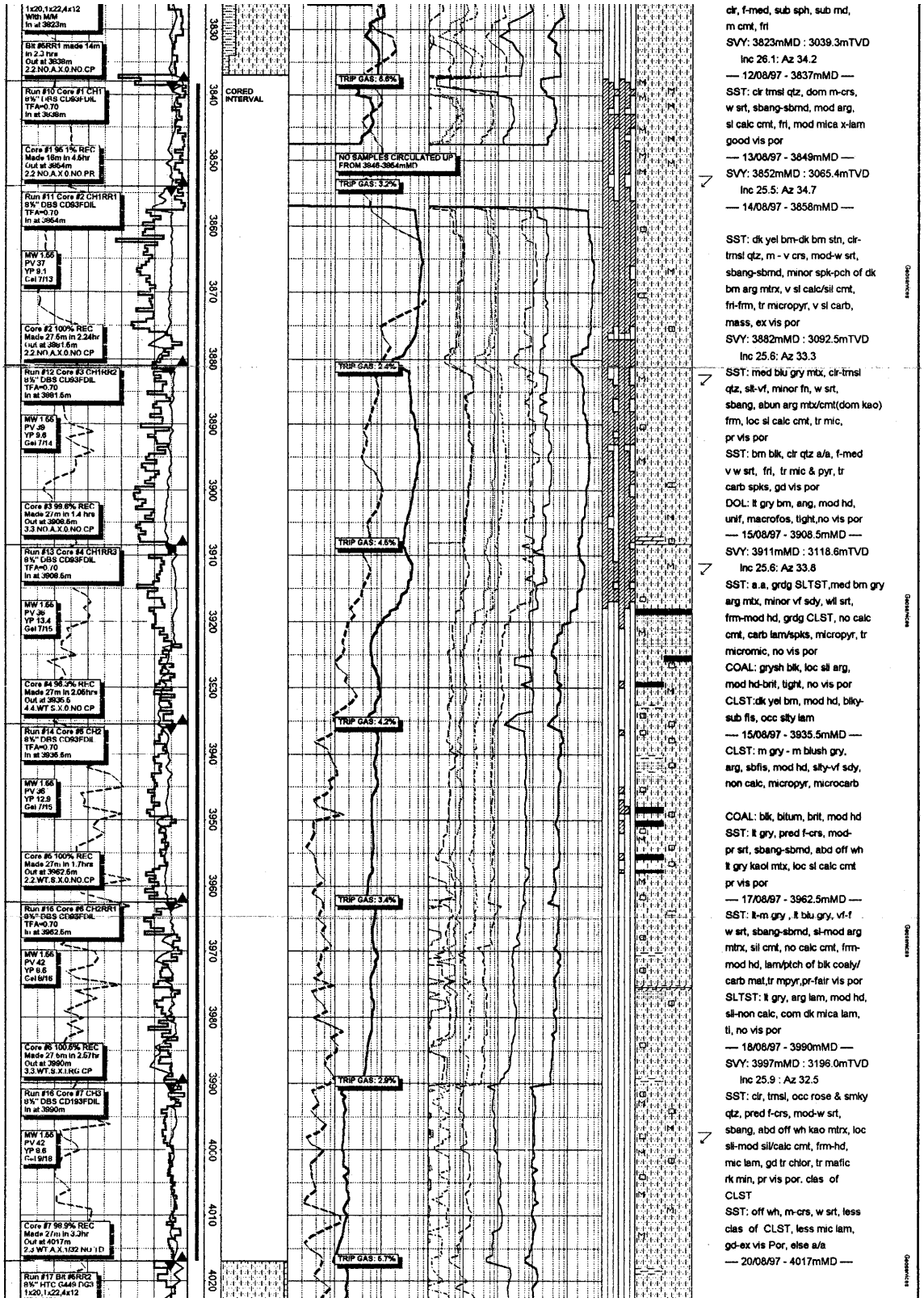


Figure 3.2 Segment of formation evaluation log from Well 15/9-19 A.

3.4. Results and Discussion

In order for the network to understand the specified lithology targets, the lithology code would be separated into 5 columns before being input into the system as a target variable. This would then give us a matrix of size 557x5, representing static data: 557 samples of 5 elements. Since three variables were used as the input (GR, RHOB, and NPHI), the input matrix would be of size 557x3, representing static data: 557 samples of 3 elements.

The dataset was then randomly divided into segments for training, validation, and testing. The training segment are presented to the network during the training phase, and the network is then adjusted according to its error. The validation segment is used to measure the ability of the network to generalize based on what it has trained upon in the previous step. The network would halt training when generalization stops improving. The testing segment has no effect on training and so provides and independent measure of network performance during and after training.

In this example, the training set consisted of 70% of the dataset (389 samples), whereas the other two sets consisted of 15% of the dataset (84 samples), respectively.

10 neurons were used in the hidden layer of the neural network. The network was trained by the use of scaled conjugate gradient backpropagation (trainscg) technique. The training would automatically stop when generalization stops improving. The generalization is tracked by use of the cross-entropy error. Lower values of the cross-entropy error are better and refer to better classification. A value of zero means that there was no error in classifying the inputs with their targets. Percentage error was also used to monitor the training of the network. This parameter indicates the percentage of samples which were misclassified. A value of 0 would indicate no misclassifications, whereas a value of 100 would indicate maximum misclassifications.

Results of the analysis yielded the following results:

Table 3.4 Results of analysis.

	Samples	*CE*	*%E*
Training	389	1.16181e-0	38.56041e-0
Validation	84	3.00524e-0	42.85714e-0
Testing	84	3.02733e-0	32.14285e-0

Varying the number of neurons in the network or the number of times the network was trained, did not lead to much improvement in the results that were obtained in terms of minimizing the errors obtained. From the above results several analysis plots may be plotted, and the MATLAB® script that was utilized may be extracted from the application's GUI.

Figure 3.3 shows the training window that is displayed during the training process. The user may use this window to track the training progress and abort it if necessary. From the training window we see that we are feeding 3 inputs (i.e. GR, RHOB, and NPHI) into the network which has 10 neurons in its hidden layer. The network then produces fives outputs

which are our lithology classes. We also note that the network is dividing the data randomly into the training, validation and testing datasets according to the previous percentages that were supplied. Training is executed by use of the scaled conjugate gradient, and the performance of the training is gauged by minimizing the cross-entropy parameter. In this case, the training stopped when the validation error started increasing for 6 consecutive iterations. This event occurred when the network reached 34 iterations. Note that the network is set to automatically terminate at 1000 iterations in case the network cannot converge to a solution.

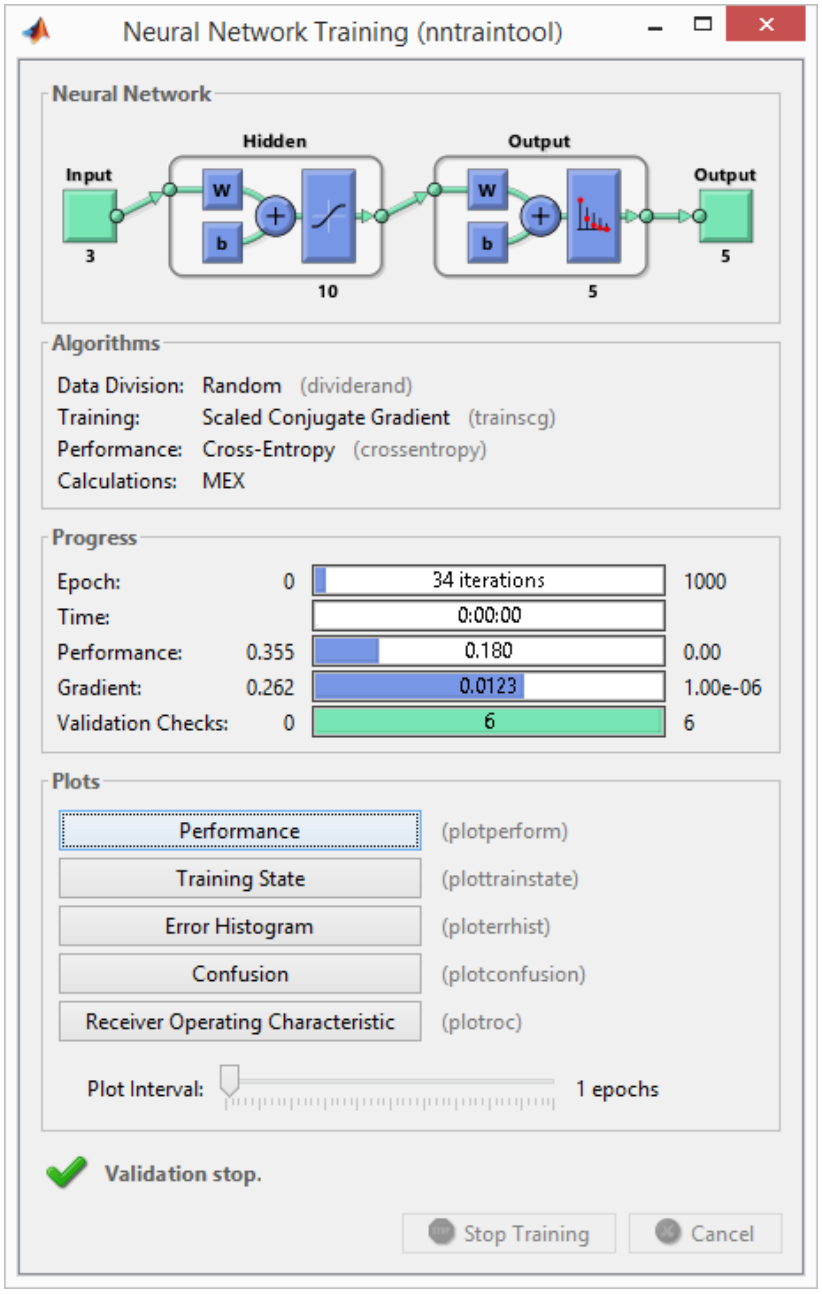


Figure 3.3 Training of the Neural Network. 34 iterations were performed until the best solution was reached.

To analyze the validation performance throughout the training process, the user may click on the *Performance* button in the training window (***Figure 3.3***). ***Figure 3.4*** displays the performance curve as a function of the training iterations. As can be observed from the curve, the cross-entropy of all three data sets drops gradually from 0 to 28 iterations/epochs. Note that the cross-entropy for the validation curve begins increasing after 28 epochs. The network was set to terminate the training process if the validation error increased for 6 validation checks in a row. The network is then restored to the point where the lowest validation error was obtained. Three curves are shown on the plot since the data set was split into 70% for training, 15% for validation and 15% for testing. Training on the training vectors will continue as long as the training algorithm reduces the network's error (cross-entropy in this case) on the validation vectors. If training were to continue, an issue of poor generalizing may occur. This is when the network begins memorizing the training dataset making it unable to properly predict the testing dataset. By stopping the training at the best validation error, we avoid the issue of poor generalization. The best validation performance achieved a cross-entropy of 0.21825 at the 28th iteration.

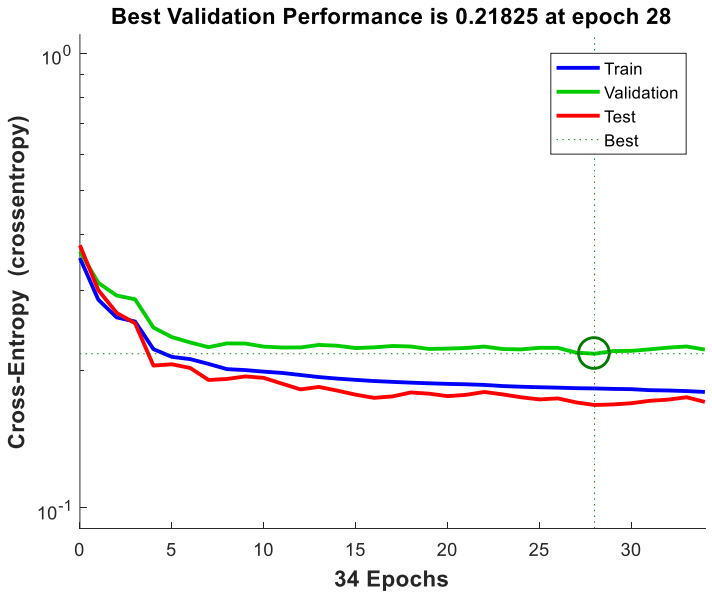


Figure 3.4 Performance plot.

The training state plot shown in ***Figure 3.5*** can be used to show us exactly where the validation checks failed. This plot is brought up by clicking on the *Training State* button on the training window (***Figure 3.3***). As can be observed from the validation check plot, the validation checks were successful from iteration 0 to 8, meaning the error kept decreasing in comparison to the previous iteration. The error began increasing from 8 to 10 iterations for 3 consecutive validation checks. Similarly the error increased from:

- 12 to 14 iterations for 3 consecutive validation checks,
- 16 to 18 iterations for 3 consecutive validation checks,
- 20 to 22 iterations for 3 consecutive validation checks,
- 25 to 26 iterations for 2 consecutive validation checks,

- and, finally, 29 to 34 iterations for 6 consecutive validation checks.

The training stopped at this point, since the network reached 6 validation failures.

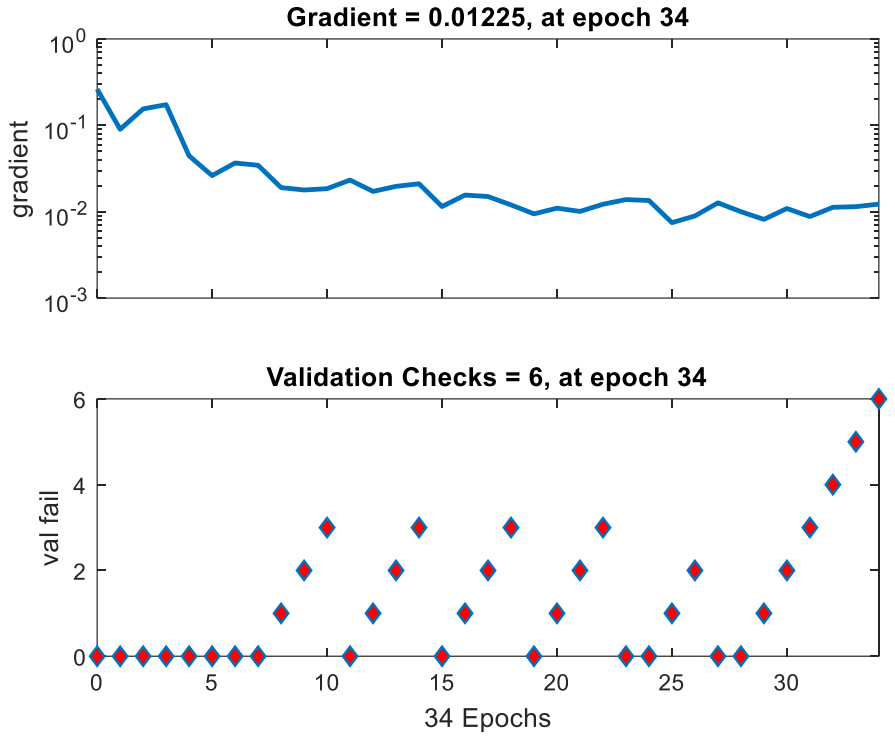


Figure 3.5 Training state plot.

The error histogram, as shown in ***Figure 3.6***, is used to display the error in prediction vs. actual results for the algorithm. The error histogram can be viewed by clicking on the *Error Histogram* button on the training window (***Figure 3.3***). A good algorithm will have errors that are very close to zero which would be represented by a peak in the middle of the error histogram. The zero error line represents the point where most of the instances should cluster around for a good network performance. In this case, the error histogram is divided into 20 bins to view where the instances would cluster. The maximum errors obtained were -0.8706 and 0.9516, however very few instances were present in these bins. The vast majority of the instances clustered around the zero error line in a bin of value -0.00744. This shows that the network performed well in minimizing the error between the targets and outputs. There is still room for improvement, however, since we can also observe that

there are also several instances which don't have optimum error achieved in the 2 bins to the left of the zero error line (i.e. bin -0.1992 and -0.1033).

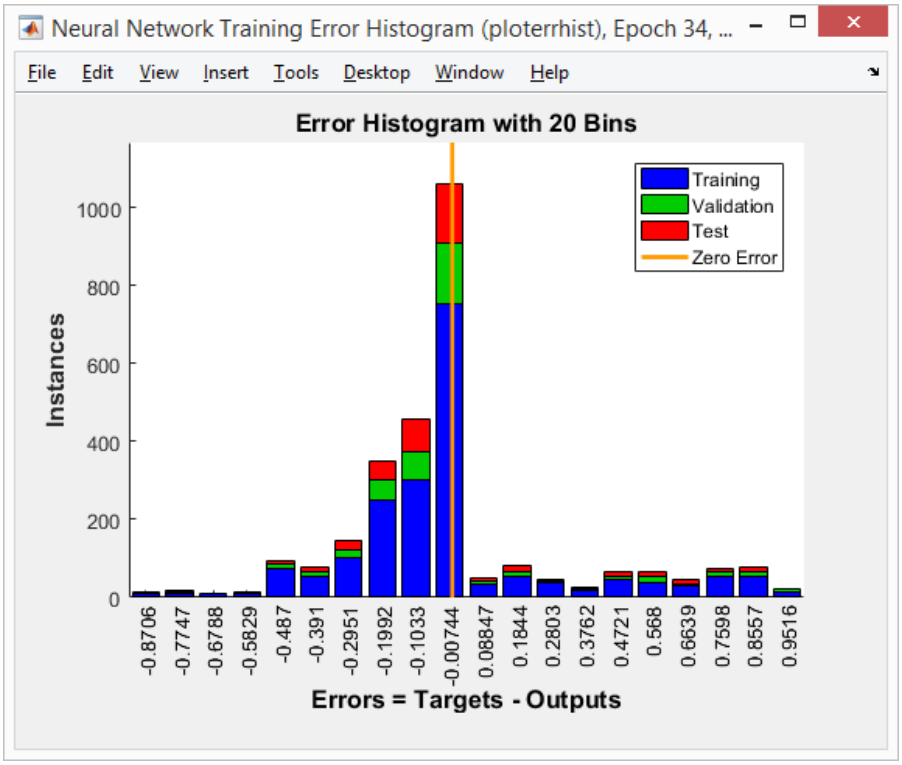


Figure 3.6 Error histogram plot.

In the confusion matrices shown in ***Figure 3.7***, the diagonal cells in each table show the number of cases that were correctly classified, and the off-diagonal cells show the misclassified cases. The blue cell in the bottom right shows the total percent of correctly classified cases (in green) and misclassified cases (in red) (The MathWorks, Inc., 2009). The individual classification results for the training, validation and testing data sets can be viewed by looking at their respective matrices. The *All Confusion Matrix* is the summation of the classifications from all of these matrices. As we can see from the plots, the network performs a good job at correctly classifying sandstone instances in the datasets (Target Class 1), achieving a positive classification percentage of 94.8%. The network performed poorly, however, in the classification of other target lithology (i.e. dolomite, clastic, coal, and siltstone). For dolomite and clastic the network achieved a positive classification percentage of 21.6% and 35.7%. For coal and siltstone the network was not able to correctly classify any of those instances. These results, however, can be explained by looking at the distribution of the lithology in the input dataset. In the input data set there were 325 instances of sandstone, 51 instances of dolomite, 70 instances of clastic, 56 instances of coal, and 55 instances of siltstone. Since the dataset is dominated by instances of Sandstone, the network tends to over predict the occurrence of sandstone. This problem can be solved by increasing the number of relevant input values that would allow for the network to generalize better. It may also be improved by obtaining a dataset that has a more balanced distribution of instances corresponding to each lithology type.

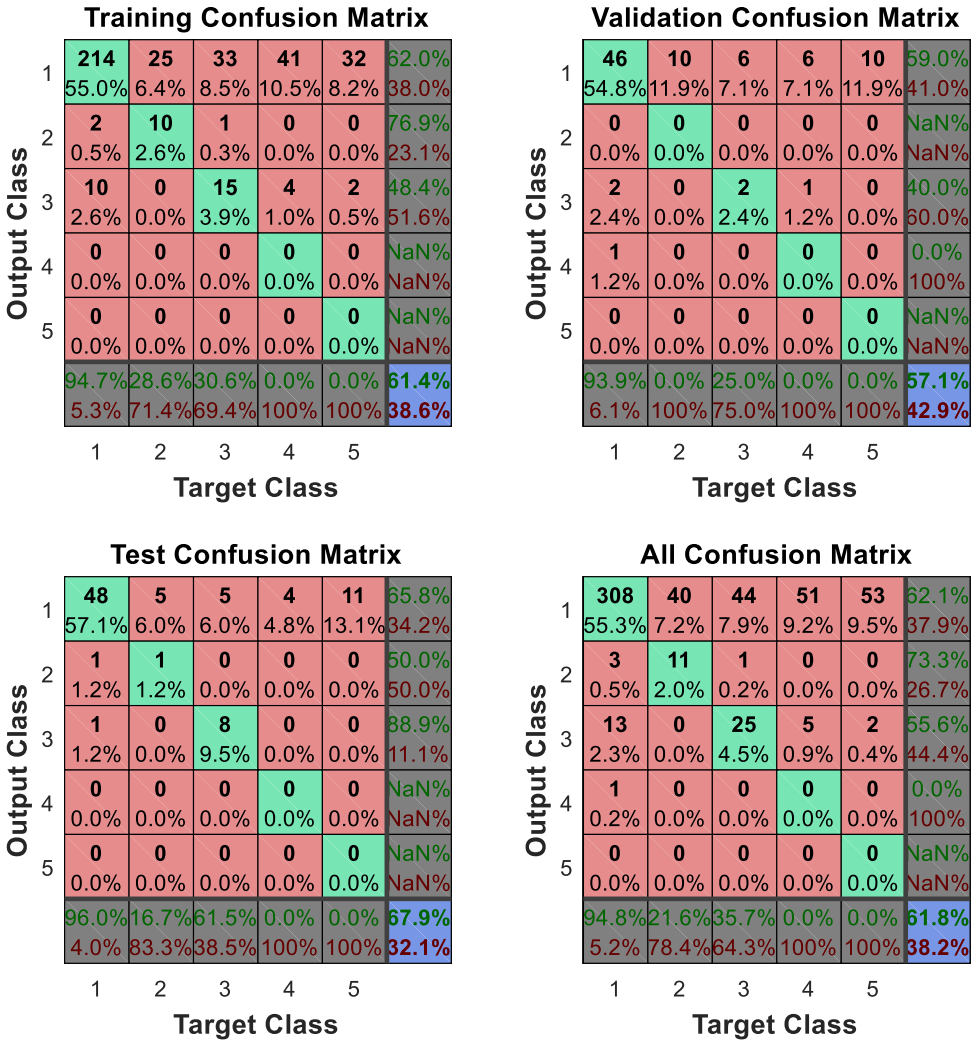


Figure 3.7 Confusion matrices.

Another way of visualizing the performance of the classifier is by the use of a Receiver Operating Characteristic curve or ROC curve. Similar to how the confusion matrix works, the ROC curve is used to distinguish between how well the classifier classifies true positives (TP) and true negatives (TN). The curve is constructed by plotting the true positive rate vs. the false positive rate at different threshold settings as can be observed in ***Figure 3.8***. Due to this comparison between true positive rate and false positive rate, the ROC curve is also sometimes referred to as the *relative operating characteristic* curve. Lastly, the ROC curve can also be viewed as a display of the sensitivity (probability of predicting a real positive as a positive) vs. the 1 – sensitivity (probability of predicting a real negative as a positive) (What is an ROC curve?, n.d.). A perfect test would show points

in the upper-left corner, with 100% sensitivity and 100% specificity (The MathWorks, Inc., 2009).

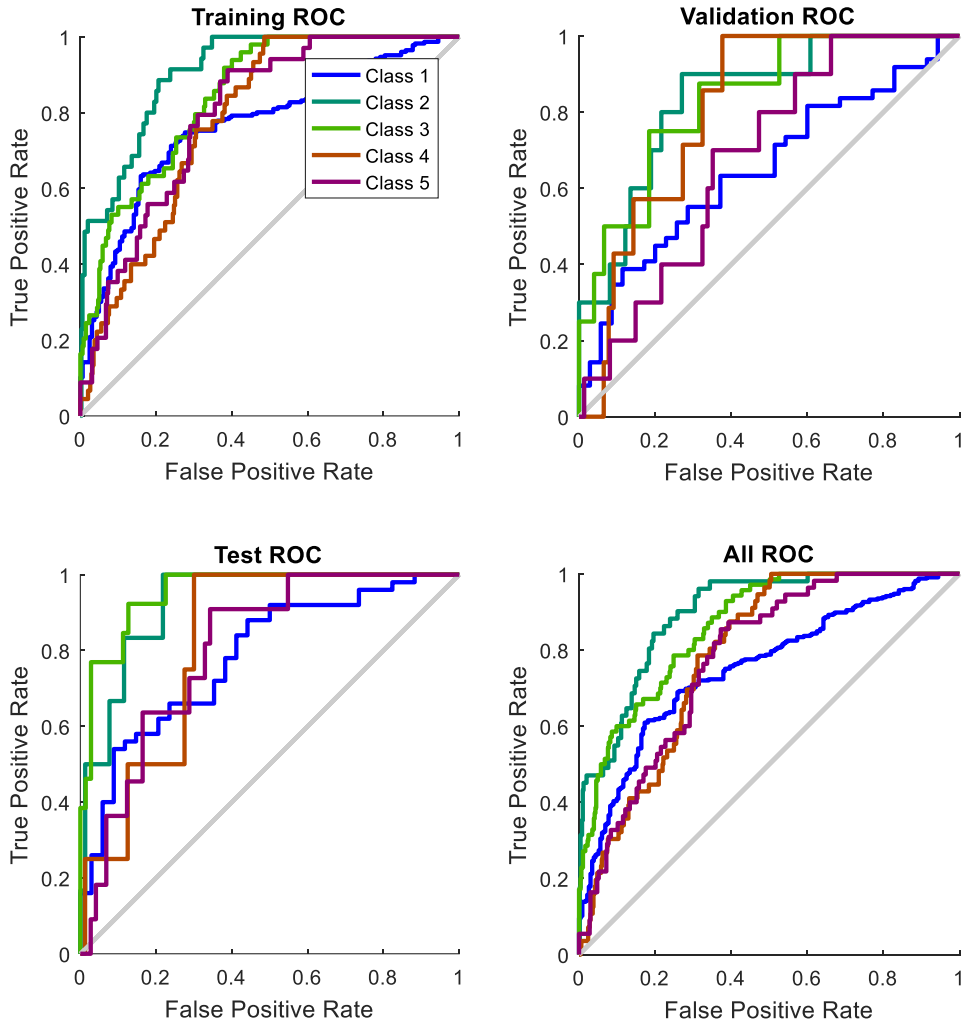


Figure 3.8 Receiver Operating Characteristic (ROC) plot.

Table 3.5 MATLAB code extracted from the GUI.

```
% Solve a Pattern Recognition Problem with a Neural Network
% Script generated by Neural Pattern Recognition app
% Created 19-Apr-2019 09:22:53
%
% This script assumes these variables are defined:
%
%   Inputs - input data.
%   Targets - target data.

x = Inputs';
```

```
t = Targets';

% Choose a Training Function
% For a list of all training functions type: help nntrain
% 'trainlm' is usually fastest.
% 'trainbr' takes longer but may be better for challenging problems.
% 'trainscg' uses less memory. Suitable in low memory situations.
trainFcn = 'trainscg';  % Scaled conjugate gradient backpropagation.

% Create a Pattern Recognition Network
hiddenLayerSize = 10;
net = patternnet(hiddenLayerSize, trainFcn);

% Setup Division of Data for Training, Validation, Testing
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;

% Train the Network
[net,tr] = train(net,x,t);

% Test the Network
y = net(x);
e = gsubtract(t,y);
performance = perform(net,t,y)
tind = vec2ind(t);
yind = vec2ind(y);
percentErrors = sum(tind ~= yind)/numel(tind);

% View the Network
view(net)

% Plots
% Uncomment these lines to enable various plots.
%figure, plotperform(tr)
%figure, plottrainstate(tr)
%figure, ploterrhist(e)
%figure, plotconfusion(t,y)
%figure, plotroc(t,y)
```

3.5. Conclusions

- As can be observed from the confusion matrix in ***Figure 3.7***, there is still plenty of room for improvement. By use of the parameters specified in the methodology, an overall classification percentage of 61.8% was achieved, whereas 38.2% of the targets were misclassified. As previously stated, changing the number of neurons in the network and the initial training parameters or the number of times the network was trained, had little effect on the final results. Note, however, that the sample size used in this experiment (557 samples) may have been too small for the network to generalize sufficiently. The use of a larger sample size may perhaps help the network to generalize better and hence yield better results.

- The ROC curve in ***Figure 3.8*** shows that for this particular dataset, Class 2 (Dolomite) achieved, on average, a higher true positive rate in comparison to the other lithology classes. Moreover, Class 1 (Sandstone) achieved, on average, a lower true positive rate in comparison to the remaining lithology classes.
- The error histogram in ***Figure 3.6*** illustrates that the algorithm performs well in minimizing the errors between predicted and actual values. The vast majority of the instances fall in the bin with an error value very close to zero (-0.00744) for all stages of developing the model (i.e. training, validation and testing).

To sum up, the use of ANN in the classification of lithology is important to the determination of suitable CO_2 sequestration sites since the lithology plays a major role in the rock properties (such as permeability and porosity) that would be present. These properties are important in order to permit the movement of CO_2 into the pore space once injected into the formation via injection wells. The formation would also need to have sufficient porosity to allow the storage of acceptable quantities of CO_2. It is an added advantage to governments and institutions seeking to store CO_2 to not only be able to predict individual rock properties, but also a generic characteristic such as lithology to aid in the finding of these reservoirs. The more data that can be provided to the network and the more relevant the data is to the problem at hand, the better we'll be able to locate these reservoirs.

3.6. Acknowledgements

The author would like to thank his thesis advisory committee for the feedback and suggestions that aided in the enhancement of this chapter.

3.7. References

Al-Anazi, A., & Gates, I. D. (2010). On the Capability of Support Vector Machines to Classify Lithology from Well Logs. *Natural Resources Research*, 125-139.

Benaouda, D., Wadge, G., Whitmarsh, R. B., Rothwell, R. G., & MacLeod, C. (1999). Inferring the lithology of borehole rocks by applying neural network classifiers to downhole logs: an example from the Ocean Drilling Program. *Geophysical Journal International*, 477-491.

Bhatt, A., & Helle, H. B. (2002). Determination of facies from well logs using modular neural networks. *Petroleum Geoscience*, 217-228.

Busch, J. M., Fortney, W. G., & Berry, L. N. (1987). Determination of Lithology From Well Logs by Statistical Analysis. *SPE Formation Evaluation.* Society of Petroleum Engineers.

Deng, C., Pan, H., Fang, S., Konaté, A. A., & Qin, R. (2017). Support vector machine as an alternative method for lithology classification of crystalline rocks. *Journal of Geophysics and Engineering*, 341-349.

Ghosh, S., Chatterjee, R., & Shanker, P. (2016). Estimation of ash,

moisture content and detection of coal lithofacies from well logs using regression and artificial neural network modelling. *Fuel*, 279-287.

Imamverdiyev, Y., & Sukhostat, L. (2019). Lithological facies classification using deep convolutional neural network. *Journal of Petroleum Science and Engineering*, 216-228.

Kapur, L., Lake, L. W., Sepehrnoori, K., Herrick, D. C., & Kalkomey, C. T. (1998). Facies Prediction From Core And Log Data Using Artificial Neural Network Technology. *SPWLA 39th Annual Logging Symposium, 26-28 May*. Keystone, Colorado: Society of Petrophysicists and Well-Log Analysts.

Lewis, H. G., & Brown, M. (2001). A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, 3223-3235.

Ma, Y. Z. (2019). Facies and Lithofacies Classifications from Well Logs. In Y. Z. Ma, *Quantitative Geosciences: Data Analytics, Geostatistics, Reservoir Characterization and Modeling* (pp. 231-254). Springer, Cham.

Mahmoodi, O., Smith, R. S., & Tinkham, D. K. (2016). Supervised classification of down-hole physical properties measurements using neural network to predict the lithology. *Journal of Applied Geophysics*, 17-26.

Ojha, M., & Maiti, S. (2016). Sediment classification using neural networks: An example from the site-U1344A of IODP Expedition 323 in the Bering Sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, 202-213.

Saggaf, M. M., & Nebrija, E. L. (2000). Estimation of Lithologies and Depositional Facies from Wire-Line Logs. *AAPG Bulletin*, 1633-1646.

Sammut, C., & Webb, G. I. (2017). *Encyclopedia of Machine Learning and Data Mining*. New York: Springer Science. doi:10.1007/978-1-4899-7687-1

Silva, A. A., Neto, I. A., Misságia, R. M., Ceia, M. A., Carrasquilla, A. G., & Archilha, N. L. (2015). Artificial neural networks to support petrographic classification of carbonate-siliciclastic rocks using well logs and textural information. *Journal of Applied Geophysics*, 118-125.

Son, S., Hou, J., Liu, Y., Cao, S., Hu, C., Wang, X., & Chang, Z. (2016). Application of artificial neural network in Geology: Porosity estimation and lithological facies classification. *12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. Changsha, China: IEEE.

Srinivasan, S., & Petkovic, D. (2000). Phonetic confusion matrix based

spoken document retrieval. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 81-87). Athens: ACM Digital Library.

Tang, H., & White, C. D. (2008). Multivariate statistical log log-facies classification on a shallow marine reservoir. *Journal of Petroleum Science and Engineering*, 88-93.

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, 40-50.

Visa, S., Ramsay, B., Ralescu, A., & Knaap, E. v. (2011). Confusion Matrix-based Feature Selection. *Proceedings of the Twenty--second Midwest Artificial Intelligence and Cognitive Science Conference* (pp. 120-127). Cincinnati: Omnipress - Madison, WISCONSIN.

Walls, J. D., Taner, M. T., Guidish, T., Taylor, G., Dumas, D., & Derzhi, N. (1999). North Sea reservoir characterization using rock physics, seismic attributes, and neural networks; a case history. *SEG Technical Program Expanded Abstracts* (pp. 1572-1575). Society of Exploration Geophysicists.

What is an ROC curve? (n.d.). Retrieved from The Analysis Factor: https://www.theanalysisfactor.com/what-is-an-roc-curve/

4. CHAPTER 4 - POROSITY PREDICTION VIA ARTIFICIAL NEURAL NETWORKS (ANN)

4.1. Abstract

When locating a suitable CO_2 sequestration reservoir, one of the major rock properties to be looking for is the porosity of the rock. The porosity is a measure of the pore volume of the formation divided by its bulk volume. It is this property that would determine the quantity of CO_2 that can be injected into the reservoir. Due to its importance, knowledge of the porosity of a particular rock formation is crucial in the making of many decisions. It is quite costly and time consuming to obtain the accurate porosity data. ANN can be used to aid in the prediction of porosity in wellbores by use of available porosity data from offset wells. Several authors have investigated the prediction of porosity by use of AI techniques. The techniques mainly make use of either core, log, or seismic data. It is also common to use a combination of these.

In this study, we use well log data obtained from a well in the Volve field, present in the Norwegian Continental Shelf, to predict porosity data. The output is compared to porosity data obtained from a routine analysis of 7 different core samples collected from the same well. Three well logs were used; gamma ray (GR), bulk density (RHOB), and neutron porosity (NPHI), to predict porosity data obtained from 664 core plugs. A single layered artificial neural network (ANN) was utilized to train 70% of the data points and test the remaining 30% of the data. The optimization of results was performed by comparing three main statistical parameters, which are the correlation coefficient (CC), the root mean-squared error (RMSE) and the absolute average percentage error (AAPE).

Results show the optimum number of neurons in the network to be 18, where, the CC was 0.808, RMSE was 0.055, and AAPE was 5.94. A mathematical expression that can be used to calculate the tan of the predicted porosity (ϕ), by the use of weights and biases extracted from the network, is also presented.

4.2. Introduction

We continue our discussion on the characterization of CO_2 sequestration sites with the prediction of porosity via artificial neural networks. This approach has already been used in a previous study that made use of the methodology to predict the permeability of prospective formations in the Volve Field, Norwegian Continental Shelf.

Knowledge of the porosity of a rock formation is crucial in the determination of possible reserves that may be present and the saturations of the various fluids that could be present in the formation. Several studies have been carried out to determine it by means of artificial intelligence. Jamshidian *et al*. (2015) utilized ANN along with Multi Linear Perceptron (MLP) to predict NMR logging parameters. Nuclear Magnetic Resonance (NMR) is becoming an increasingly popular method employed to estimate the porosity and permeability of formation layers due to its effectiveness. In the field, tools that employ NMR also have the additional advantage of not requiring radioactive sources to determine porosity. However, the use of NMR is not also possible in certain scenarios. The use of

ANN in these cases offers a solution by estimating the relevant NMR parameters from data that has already been obtained from other sections of a formation. The team enhanced their neural network by the use of the Imperialist Competitive Algorithm (ICA). This system was then deployed onto two independent data sets obtained from the south Pars gas field. Their results show that the model outperforms conventional neural network techniques. Mahmood and Ahmad (2017) made use of a multi-layer feed forward neural network (MLFN) in the estimation of porosity. The estimation of porosity may prove challenging due to the failure of comprehending the spatial porosity parameter distribution throughout the formation. The authors utilized well log and seismic data that included P-impedance, density, fluid, Vp/Vs, etc from the Badin field in Pakistan. The model that was developed was then used to predict the porosity with a correlation coefficient of 0.91. Singh, Kanli and Sevgen (2016) designed a back-propagation artificial neural network to estimate porosity in a Kansas gas field, USA. The authors utilized well log data that consisted of resistivity, sonic, and density logs to estimate the parameter. Results show favorable performance in comparison to available empirical relationships. Konaté, Pan, Khan and Yang (2015) predicted porosity by the use of 4 wells obtained from the Zhenjing oilfield. They made use of a generalized regression neural network (GRNN) and trained it by use of one of the wells to establish the model. The model was then used to predict porosity in three other wells obtained from the same field. The authors conclude that GRNN predicts porosity better than other commonly used ANN methods. An, Yang and Zhang (2018) utilized compensated neutron, acoustic time difference, natural gamma and compensated density well logs to predict porosity via deep learning technology. The authors point out that porosity is influenced by a range of geological factors that include buried depth, the position of tectonic plates, and the depositional environment that may complicate predicting it due to its naturally heterogenous nature. It is of great importance for engineers and practitioners to be able to accurately gauge the porosity of the formation for the purpose of determining fluid saturations and minimizing the overall risk of oil and gas exploration. The end result is a non-linear relationship between logging parameters and porosity that can be used to effectively estimate porosity.

Maurya and Singh (2019) predominantly made use of seismic data to infer the porosity variations along a sand channel. The methodology involved the marrying of well-log and band-limited seismic data to get high-resolution impedance volume via acoustic impedance inversion. The authors utilized two AI schemes:

- Coloured Inversion (CI)
- Model-Based Inversion (MBI)

The study area of choice was the Blackfoot region, Alberta, Canada. The methodology involves the prediction of impedance by the use of these two schemes, thereafter, the porosity volume is predicted by the use of a multivariate regression and the Probabilistic Neural Network (PNN) using the obtained impedances as inputs to the model. The authors conclude that a combination of model-based inversion and PNN can produce a more reliable estimate of formation evaluation properties such as porosity. Similarly, from a gas

field located in Indonesia, Basri, Aswad, Suryana and Priatama (2018) made use of 2D seismic data to generate acoustic impedance using an inversion based methodology. The impedance was then used in parallel with other seismic attributes as inputs into multiarribute transforms and PNN. Out of the attributes that were studied, the authors concluded that two of them used in a combination offered good results for predicting porosity. They reached a cross correlation of 72% with multiattribute transforms which then increased to 92% by use of PNN. Furthermore, from data obtained from the Upper Assam basin in northeastern India, Gogoi and Chatterjee (2019), absolute and relative acoustic impedances (AAI and RAI) were generated from 2D seismic data. Porosity was obtained by use of wavelet and reflectivity from the seismic data. These were then used as inputs to a multilayered feed forward neural network (MLFN).

Elkatatny, Tariq, Mahmoud and Abdulraheem (2018) used ANN to predict porosity by use of bulk density, neutron porosity, and sonic compressional time. The authors were able to achieve prediction results with a correlation coefficient of 0.98 and an AAPE of less than 8%. The authors point out that an added advantage of using ANN in contrast with other AI techniques such as Support Vector Machine (SVM) and Adaptive Neuro-Fuzzy Inference System (ANFIS) is the ability to extract a mathematical relationship, consisting of weights and biases, that can be used to predict the porosity without the need to re-train and test the model.

Gu *et al*. (2017) utilized an improved statistical approach to predict porosity and permeability from wells in the LULA oilfield in the Santos Basin. They propose two algorithms: (i) stepwise regression and (ii) N-way analysis of variance. Both algorithms offer better selection of well log data that would be significant in the prediction of porosity and permeability. In comparing the two algorithms, the authors concluded that the N-way analysis of variance, in general, offered more accurate results.

Konaté, Pan, Khan, and Ziggah (2015) predicted porosity from geophysical well log data by the use of a feed-forward back propagation (FFBP) neural network and radial basis function (RBF). Input data that was used included density, neutron porosity, sonic and resistivity. The data was obtained from wells drilled as part of the Chinese Continental Scientific Drilling Main Hole. The authors point out that ANN provides superior results in comparison to regression techniques in the case of three sets of well logs (density, sonic and resistivity). A comparison of FFBP and RBF also revealed that RBF outperformed FFBP in this study. Moon, Lee, Kim, Choi, and Kim (2016) carried out their study by the use of collocated cokriging (CCK) and neural-network multi-attribute transform (NN-MAT). The dataset consisted of 3D seismic data as well as well log data from a field in Wyoming, USA. Results show that CCK overpredicted porosity in the wells that it was tested on, whereas NN-MAT generally underpredicted porosity. The authors were aiming to answer two main questions: (i) how do these two methods perform in the prediction of porosity and (ii) how does the number of wells included in the dataset affect the results. Duan, Li, Li, and Sun (2016) constructed a committee neural network (CNN) that was based on various individual neural networks (i.e. back propagation, radial basis function

and support vector regression) to predict porosity. Their dataset included three well logs, where one was used to train the model and the other two were used for testing. Lastly, Mahmood, Shakir, Abuzar, Khan and Khattak (2017) used ANN to predict porosity in the Balkassar oil field by the use of 3D seismic data and well logs. They utilized the seismic data to generate acoustic impedance by inverting the data.

4.3. Methodology

Gamma Ray (GR), Bulk Density (RHOB), and Neutron Porosity (NPHI) well logs were obtained from Well 15/9-19 A located in the Volve field, Norwegian Continental Shelf. The logs, which were presented in image format, were digitized by the use of a MATLAB® code. A total of 4356 data points were generated for GR. 6912 data points and 5595 data points were generated for RHOB and NPHI, respectively. ***Figure 4.1*** shows the digitized input logs. The GR log was calibrated from 0 to 150 gAPI and 3470 to 4127 m. The RHOB log was calibrated from 1.95 to 2.95 g/cc and 3470 to 4127 m. The NPHI log was calibrated from 45 to -14 pu and 3470 to 4127 m. Care needs to be taken to ensure that the peaks of the logs are captured.

Log values were then correlated with corresponding core porosity values obtained from core analysis of 664 core plugs. These three input paramaters are then input in the network and trained by use of the known porosity values. The output is provided to the training algorithm for 70% of the data set so that the algorithm can learn the trends in the log data that would result in the specified outputs. The algorithm is then tested on the remaining 30% of the data set to gauge how well it is able to predict the porosity. The optimizatioin of the results is achieved by comparing three main statistical parameters that include: the correlation coefficient (CC), the root mean-squared error (RMSE) and the absolute average percentage error (AAPE).

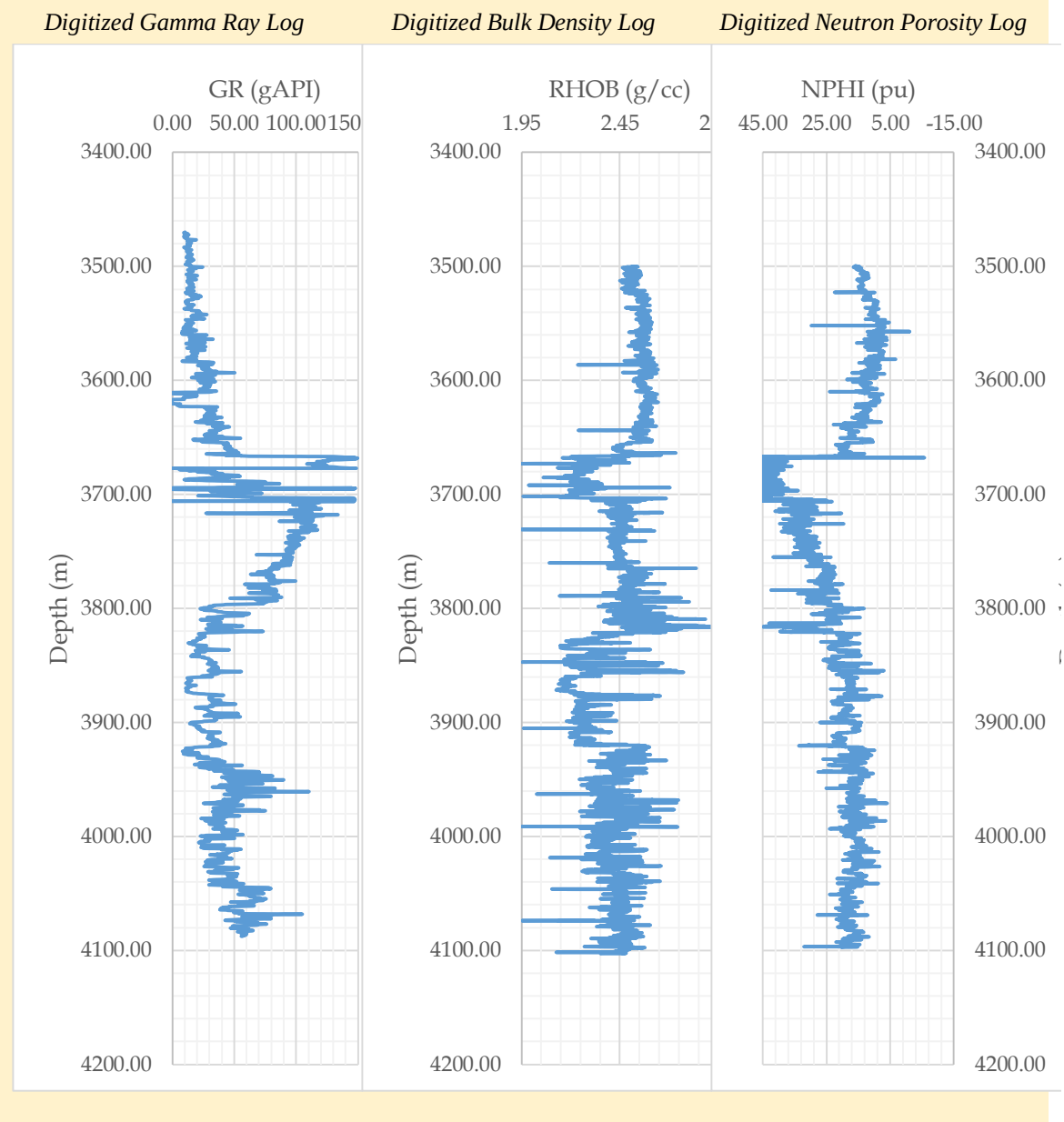


Figure 4.1 Digitized input well logs [Generated data points: GR = 4356 data points, RHOB = 6912 data points, NPHI = 5595 data points].

4.4. Results

Figure 4.2 shows the results of correlation coefficients obtained between the results of the computation vs. the core porosity for 100 different iterations. The iterations differ in the way the data is randomized. Out of the 664 core data points, 500 were initially randomly selected to train the model. 100 different random selections were conducted by use of a 'for loop' and the rand('seed') function available in MATLAB® to fix the random selection and ensure reproducibility.

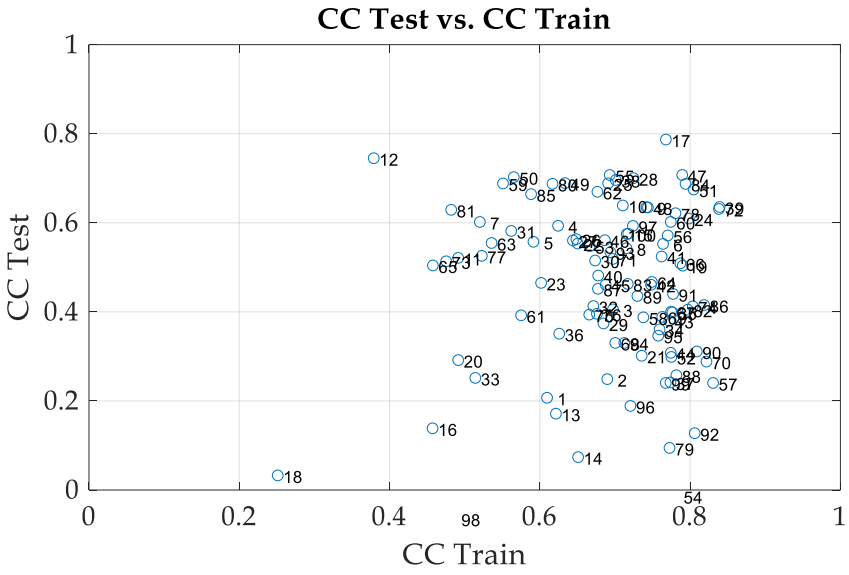


Figure 4.2 Cross-plot of the correlation coefficient between the results of the computation and the output for the case of input vs. output without taking the logarithmic value of any of the variables. The y-axis displays the CC with the tested data points (30%), and the x-axis displays the CC with the data points used for training (70%).

As can be observed from ***Figure 4.2***, the scatter of the data is quite large. These results are undesirable since they show that the performance of the ANN model is highly sensitive on the selection of the training and testing data sets.

The same analysis was performed for the case of taking the logarithm of the output values before training the model. As shown in ***Figure 4.3***, taking the logarithm of the output values slightly helps to improve correlation and reduce scatter.

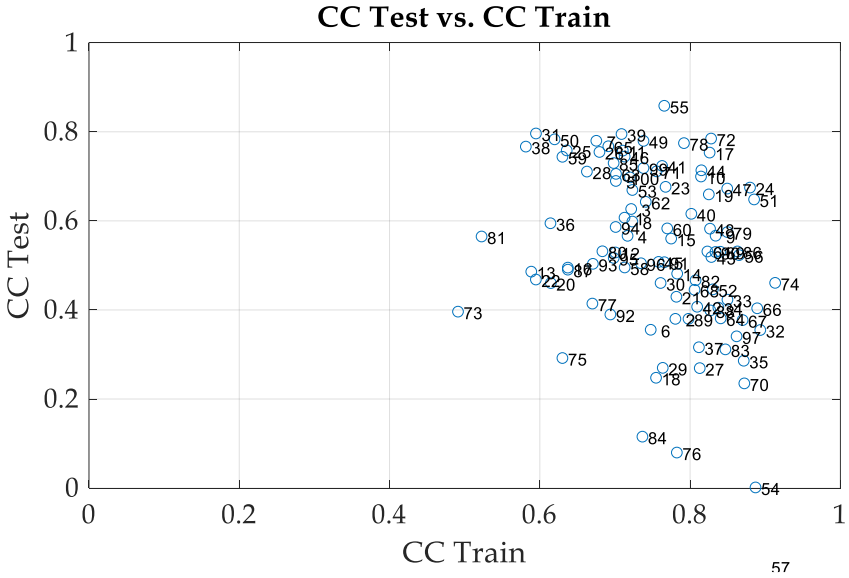


Figure 4.3 Cross-plot of the correlation coefficient between the results of the computation and the output for the case of input vs. log(output). The y-axis displays the CC with the tested data points (30%), and the x-axis displays the CC with the data points used for training (70%).

Finally, the analysis was performed for the case of taking the logarithm of both the input and output variables. ***Figure 4.4*** displays the results of this analysis. As can be observed, this help to bring the data points close, however there still exist a few outliers (point 10 and 20). The results illustrate how changing the domain of the input and output variable can help in improving the correlation between the results of the computation and the output that is being predicted.

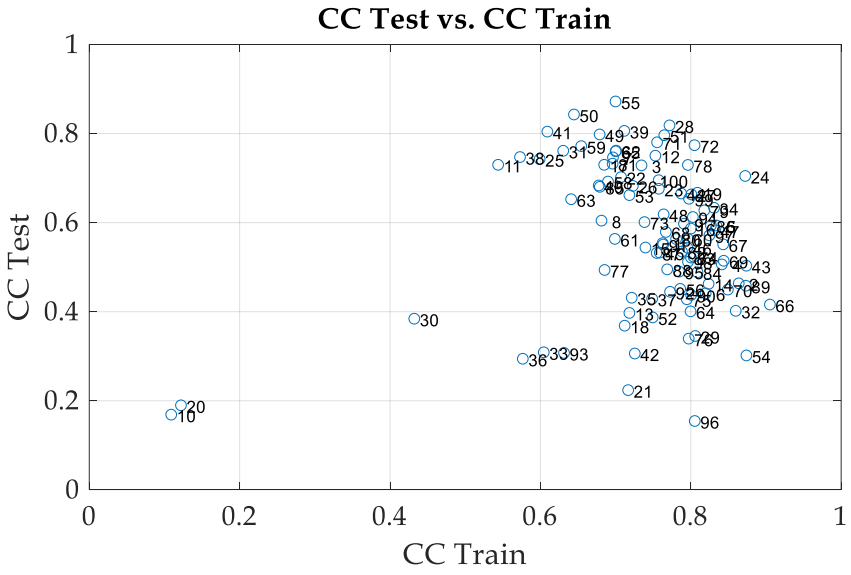


Figure 4.4 Cross-plot of the correlation coefficient between the results of the computation and the output for the case of log(input) vs. log(output). The y-axis displays the CC with the tested data points (30%), and the x-axis displays the CC with the data points used for training (70%).

Next, we conduct several different scenarios that involve varying the number of randomly selected data points from the total of 664 data points, the transfer function used to produce the output, and the number of neurons in the layer. In each case 100 iterations of different random selections were performed to determine the optimal solution. ***Table 4.1*** summarizes the results of these scenarios.

Table 4.1 Summary of the performance of different scenarios.

Case	*No. of Randomly Selected Data Points*	*No. of Neurons in Layer*	*Transfer Function*	*CC Test*	*RMSE Test*	*AAPE Test*
01	100	20	tansig	0.689	0.115	16.77
02	125	20	tansig	0.337	0.191	17.49
03	150	20	tansig	0.446	0.069	7.10
04	175	20	tansig	0.319	0.108	13.31
05	200	20	tansig	0.609	0.125	14.85
06	225	20	tansig	0.783	0.057	6.08
07	250	20	tansig	0.437	0.090	11.02
08	275	20	tansig	0.728	0.063	6.47
09	300	20	tansig	0.683	0.072	8.19
10	325	20	tansig	0.358	0.124	14.08
11	350	20	tansig	0.629	0.091	10.28
12	375	20	tansig	0.660	0.101	13.08
13	400	20	tansig	0.695	0.105	14.07

14	425	20	tansig	0.616	0.121	15.68
15	450	20	tansig	0.504	0.126	18.32
16	475	20	tansig	0.639	0.113	14.10
17	500	20	tansig	0.638	0.109	14.18
18	525	20	tansig	0.648	0.102	13.73
19	550	20	tansig	0.658	0.107	13.98
20	600	20	tansig	0.695	0.099	13.31
21	625	20	tansig	0.572	0.113	15.03
22	650	20	tansig	0.587	0.105	13.75
23	664	20	tansig	0.560	0.110	13.47
24	225	10	tansig	0.700	0.064	6.81
25	225	11	tansig	0.601	0.072	8.23
26	225	12	tansig	0.692	0.068	6.66
27	225	13	tansig	0.602	0.082	9.92
28	225	14	tansig	0.764	0.059	6.35
29	225	15	tansig	0.723	0.066	7.06
30	225	16	tansig	0.628	0.075	7.84
31	225	17	tansig	0.715	0.071	7.86
32	225	18	tansig	0.808	0.055	5.94
33	225	19	tansig	0.706	0.070	7.01
34	225	18	purelin	0.566	0.080	8.59
35	225	18	logsig	0.637	0.072	7.21

The estimated optimum number of randomly selected data points from the total of 664 was found to be 225 by holding the number of neurons in the hidden layer constant and holding the transfer function constant. The number of randomly selected data points was varied from 100 to 650 with increments of 25. A case with all of the 664 data points is also presented. The CC for the tested data was 0.783, RMSE was 0.057 and AAPE was 6.08. Next, the number of neurons was varied from 10 to 20 in increments of 1. In this case, 225 randomly selected data points were used, which is the optimum that was found from the previous analysis. The optimum number of neutrons was found to be 18. The CC was 0.808, RMSE was 0.055, and AAPE was 5.94. Finally, two other transfer functions were used that included the linear transfer function (*purelin*) and the log-sigmoidal transfer function (*logsig*). The tan-sigmoidal transfer function was found to perform better than the other two. Varying the stated parameters in these scenarios may help to slightly tweak the final results, however, the most significant differences in prediction results comes from the initial setting up of the problem. The better the correlation coefficient between the input and output variables (i.e. the more relevant the input), the better the prediction results. All in all, a total of 35 cases were performed, where the best results out of the 100 iterations would be tabulated. ***Figure 4.5*** and ***Figure 4.6*** illustrates the results of the regression analysis and the performance of the algorithm for this best-case scenario (*Case 32*). Moreover, Figure shows the tan of the predicted and original permeability as a function of depth for this scenario.

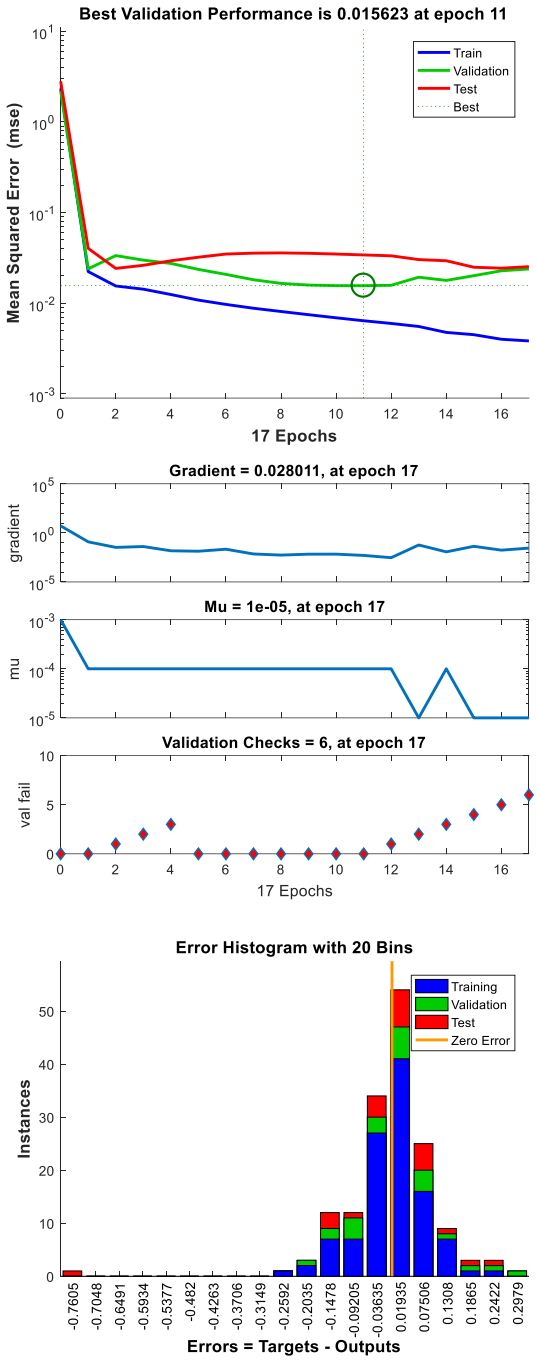


Figure 4.5 Performance of the algorithm for Case 32. The performance was gauged by attempting to minimize the mean squared error (MSE).

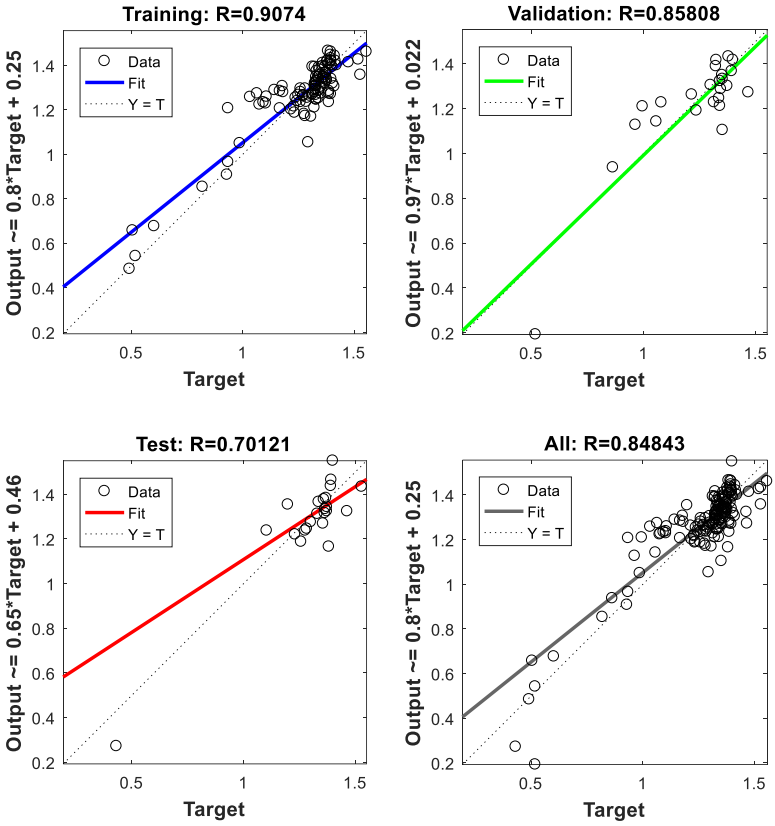


Figure 4.6 Regression analysis for Case 32 (30% Training, 15% Validation, 15% Testing).

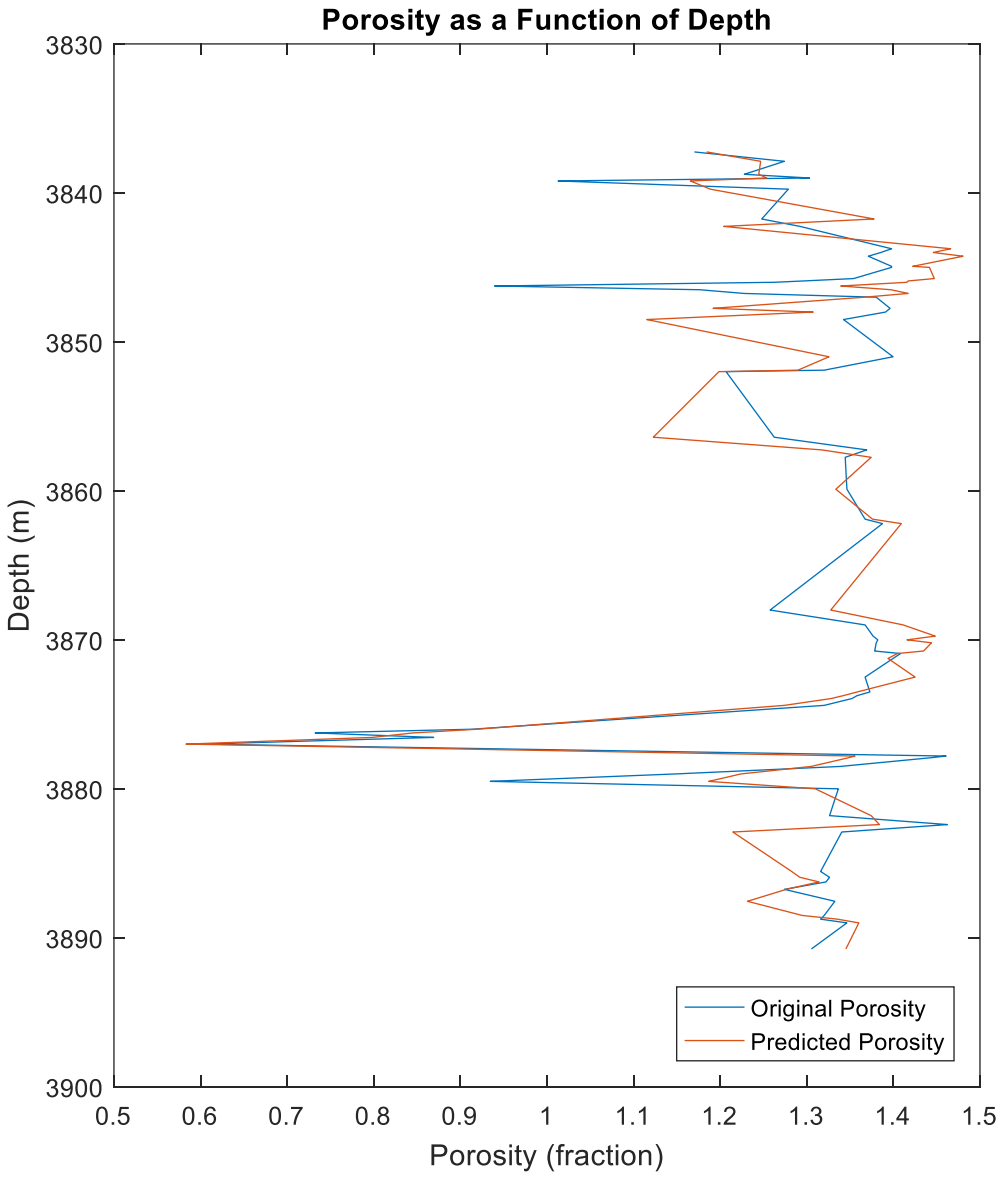


Figure 4.7 Predicted and original porosity as a function of depth for Case 32.

The weights and biases used in the developed ANN model may then be extracted from the network. ***Table 4.2*** tabulates the extracted weights for the input and hidden layer, and the extracted biases for the input and the output layer. These values may then be used to calculate the tan of the predicted porosity (ϕ) by the use of the following mathematical expression:

$$\tan(\varphi) = \left[\sum_{j=1}^{N} w_{2i}\, tansig\left(\sum_{j=1}^{J} w_{1i,j}x_j + b_{1i}\right)\right] + b_2 \qquad (4.1)$$

where $i = 1$ represents the logarithm of the gamma ray log values, $i = 2$ represents the logarithm of the bulk density log values, $i = 3$ represents the logarithm of the neutron porosity log values, N is the number of neurons (18 in this case), and J is the number of input variables (3 in this case).

Table 4.2 Extracted weights and biases from the developed ANN model.

Input Layer Weight Matrix			*Input Layer Bias Vector*	*Hidden Layer Weight Vector*	*Output Layer Bias Vector*
w_{11}	w_{12}	w_{13}	b_1	w_2	b_2
2.4779	1.6813	-3.0101	-4.3480	-1.5512	-1.1375
-2.0911	2.2535	1.8117	3.6098	-0.6854	
-2.1732	2.4294	-1.7082	2.7893	-0.1665	
0.0314	3.9319	-0.7798	-1.2977	-0.7526	
0.1158	2.3459	-3.0791	1.2751	-0.1008	
-0.6975	3.6704	1.5659	1.0381	-0.3038	
-2.6800	2.1925	-2.1498	0.9926	-0.7377	
-2.9410	0.9277	-0.7992	1.0131	1.5877	
-2.8548	2.0986	1.7444	0.3579	-0.4373	
-2.8888	1.0559	-2.6054	0.3551	0.8373	
2.3185	1.2293	2.1560	0.0009	1.0790	
2.7485	-2.3281	-0.1080	1.4577	-0.4808	
0.6307	2.0436	-2.8315	-1.8135	0.6785	
0.6568	3.2857	0.2999	1.9948	0.1607	
-0.5404	1.9893	-3.0336	-2.3507	-0.2522	
3.2164	-1.2764	0.0210	3.0799	0.0374	
-1.3217	-2.7181	2.0605	-3.2818	0.1680	
0.1103	-2.4600	-2.9771	3.4756	0.6244	

4.5. Conclusion

To conclude, the work in this chapter attempted to predict porosity values from well log data by use of a single layered ANN. The objective of the study was to locate a suitable CO2 sequestration site by the prediction of this rock property. Three input logs were used that were: Gamma Ray (GR), Bulk Density (RHOB), and Neutron Porosity. Log values were correlated with corresponding core porosity values obtained from core analysis of 664 core plugs. The dataset is split into two sets: 70% Training, and 30% Testing. The optimizatioin of the results is achieved by comparing three main statistical parameters that include: the correlation coefficient (CC), the root mean-squared error (RMSE) and the absolute average percentage error (AAPE). The optimum number of neutrons was found to be 18, where the tan-sigmoid transfer function was used in the network. The CC was 0.808, RMSE was 0.055, and AAPE was 5.94.

A reasonable cut-off value for porosity for CO_2 sequestration is 5%. A region or interval with a porosity greater than 5% would provide suitable storage capacity for storing CO_2. This cut-off value, however, will ultimately be decided by the party intending to sequester CO_2.

The results illustrate that ANN may be used as a tool in the prediction of porosity in unexplored areas for the sake of characterizing a CO_2 sequestration site. The porosity would play a major factor in the determination of how much CO_2 may be stored in a particular site.

4.6. Acknowledgements

The author would like to thank his thesis advisory committee for the feedback and suggestions that aided in the enhancement of this chapter.

4.7. References

An, P., Yang, X., & Zhang, M. (2018). Porosity prediction and application with multiwell-logging curves based on deep neural network. *SEG Technical Program Expanded Abstracts*, 819-823.

Basri, I., Aswad, S., Suryana, & Priatama, I. N. (2018). Porosity Prediction using Multiattribute Transforms and Probabilistic Neural Networks Analysis from Limestone Formation in BSJ Gas Field, Central Sulawesi, Indonesia. *Indonesian Petroleum Association, Forty-Second Annual Convention & Exhibition.* Indonesian Petroleum Association.

Duan, Y., Li, Y., Li, G., & Sun, Q. (2016). A New Neural Network Model for Rock Porosity Prediction. *2016 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI).* Beijing, China: IEEE.

Elkatatny, S., Tariq, Z., Mahmoud, M., & Abdulraheem, A. (2018). New insights into porosity determination using artificial intelligence techniques for carbonate reservoirs. *Petroleum*, 408-418.

Gogoi, T., & Chatterjee, R. (2019). Estimation of petrophysical parameters using seismic inversion and neural network modeling in Upper Assam basin, India. *Geoscience Frontiers*, 1113-1124.

Gu, Y., Bao, Z., Lin, Y., Qin, Z., Lu, J., & Wang, H. (2017). The porosity and permeability prediction methods for carbonate reservoirs with extremely limited logging data: Stepwise regression vs. N-way analysis of variance. *Journal of Natural Gas Science and Engineering*, 99-119.

Jamshidian, M., Hadian, M., Zadeh, M. M., Kazempoor, Z., Bazargan, P., & Salehi, H. (2015). Prediction of free flowing porosity and permeability based on conventional well logging data using artificial neural networks optimized by Imperialist competitive algorithm – A case study in the South Pars gas field. *Journal of Natural Gas Science and Engineering*, 89-98.

Konaté, A. A., Pan, H., Khan, N., & Yang, J. H. (2015). Generalized

regression and feed-forward back propagation neural networks in modelling porosity from geophysical well logs. *Journal of Petroleum Exploration and Production Technology*, 157-166.

Konaté, A. A., Pan, H., Khan, N., & Ziggah, Y. Y. (2015). Prediction of porosity in crystalline rocks using artificial neural networks: An example from the Chinese Continental Scientific Drilling Main hole. *Studia Geophysica et Geodaetica*, 113-136.

Mahmood, M. F., & Ahmad, Z. (2017). Application of Multi-Layer Feed Forward Neural Network (MLFNN) for the Prediction of Porosity: A Case Study from Lower Indus Basin, Pakistan. *The Nucleus*, 10-15.

Maurya, S. P., & Singh, K. H. (2019). Predicting Porosity by Multivariate Regression and Probabilistic Neural Network using Model-based and Coloured Inversion as External Attributes: A Quantitative Comparison. *Journal of the Geological Society of India*, 207-212.

Moon, S., Lee, G. H., Kim, H., Choi, Y., & Kim, H.-J. (2016). Collocated cokriging and neural-network multi-attribute transform in the prediction of effective porosity: A comparative case study for the Second Wall Creek Sand of the Teapot Dome field, Wyoming, USA. *Journal of Applied Geophysics*, 69-83.

Rafik, B., & Kamel, B. (2017). Prediction of permeability and porosity from well log data using the nonparametric regression with multivariate analysis and neural network, Hassi R'Mel Field, Algeria. *Egyptian Journal of Petroleum*, 763-778.

Singh, S., Kanli, A. I., & Sevgen, S. (2016). A general approach for porosity estimation using artificial neural network method: a case study from Kansas gas field. *Studia Geophysica et Geodaetica*, 130-140.

5. CHAPTER 5 – CONCLUSIONS AND FUTURE WORK

5.1. Conclusions

The objective of this study was to provide practitioners with a general workflow on how to locate suitable CO_2 sequestration sites by making use of the massive amount of data already available from the Upstream Oil & Gas Industry alongside the use of artificial intelligence techniques such as artificial neural networks.

The work in *Chapter 2* attempted to predict formation permeability obtained from core plugs in an exploratory well drilled in the Volve field located in the Norwegian continental shelf via the use of a single layered ANN model. The input data that was fed into the network were three common logs that include the gamma ray (GR) log, the bulk density (RHOB) log, and the neutron porosity (NPHI) log. It has been shown that taking the logarithm of both the input data and the output data, in this case, would improve the results of the prediction. An attempt at normalizing the input data to range between -1 and 1 using a two-point slope did not improve results of the prediction. The optimum single layered ANN model was determined to be one with 18 neurons in the hidden layer, and a log-sigmoidal transfer function. The optimum number of randomly selected data points from the total of 557 data points was found to be 225. An important observation to be made is that the initial setting up of the problem is more significant than varying parameters in the network, such as the number of neurons in the hidden layer, and the transfer function used in the hidden layer, in achieving good prediction results. This presents a systematic methodology to predict core formation permeability from well log data using ANN for the purposes of characterizing a potential CO_2 sequestration site.

Data obtained from lithology descriptions in logs of the same well were then used in a separate study aiming to classify lithology in *Chapter 3*. As can be observed from the confusion matrix in ***Figure 3.7***, there is still plenty of room for improvement. An overall classification percentage of 61.8% was achieved, whereas 38.2% of the targets were misclassified. Changing the number of neurons in the network and the initial training parameters or the number of times the network was trained, had little effect on the final results. Note, however, that the sample size used in this experiment (557 samples) may have been too small for the network to generalize sufficiently. The use of a larger sample size may perhaps help the network to generalize better and hence yield better results.

The ROC curve in ***Figure 3.8*** shows that for this particular dataset, Class 2 (Dolomite) achieved, on average, a higher true positive rate in comparison to the other lithology classes. Moreover, Class 1 (Sandstone) achieved, on average, a lower true positive rate in comparison to the remaining lithology classes.

The error histogram in ***Figure 3.6*** illustrates that the algorithm performs well in minimizing the errors between predicted and actual values. The vast majority of the instances fall in the bin with an error value very close to zero (-0.00744) for all stages of developing the model (i.e. training, validation and testing).

The use of ANN in the classification of lithology is important to the determination of suitable CO_2 sequestration sites since the lithology plays a major role in the rock properties (such as permeability and porosity) that would be present. These properties are important in order to permit the movement of CO_2 into the pore space once injected into the formation via injection wells. The formation would also need to have sufficient porosity to allow the storage of acceptable quantities of CO_2. It is an added advantage to governments and institutions seeking to store CO_2 to not only be able to predict individual rock properties, but also a generic characteristic such as lithology to aid in the finding of these reservoirs. The more data that can be provided to the network and the more relevant the data is to the problem at hand, the better we'll be able to locate these reservoirs.

The work in *Chapter 4* attempted to predict porosity values from well log data by use of a single layered ANN. The objective of the study was to locate a suitable CO_2 sequestration site by the prediction of this rock property. Three input logs were used that were: Gamma Ray (GR), Bulk Density (RHOB), and Neutron Porosity. Log values were correlated with corresponding core porosity values obtained from core analysis of 664 core plugs. The dataset is split into two sets: 70% Training, and 30% Testing. The optimizatioin of the results is achieved by comparing three main statistical parameters that include: the correlation coefficient (CC), the root mean-squared error (RMSE) and the absolute average percentage error (AAPE). The optimum number of neutrons was found to be 18, where the tan-sigmoid transfer function was used in the network. The CC was 0.808, RMSE was 0.055, and AAPE was 5.94.

The results illustrate that ANN may be used as a tool in the prediction of porosity in unexplored areas for the sake of characterizing a CO_2 sequestration site. The porosity would play a major factor in the determination of how much CO_2 may be stored in a particular site.

All in all, cut-off values for the sequestration of CO_2 may be set as shown in ***Figure 5.1***. By use of these cut-off values, the locations of possible sites may be highlighted as shown in ***Figure 5.2***.

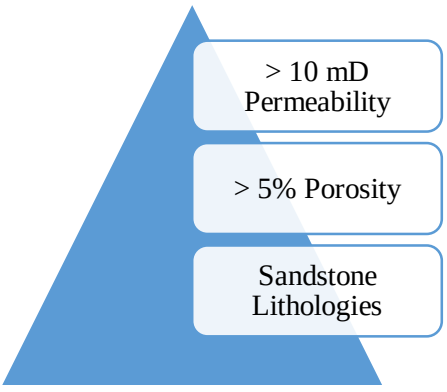


Figure 5.1 Cut-off Values for Favourable CO_2 Sequestration Locations.

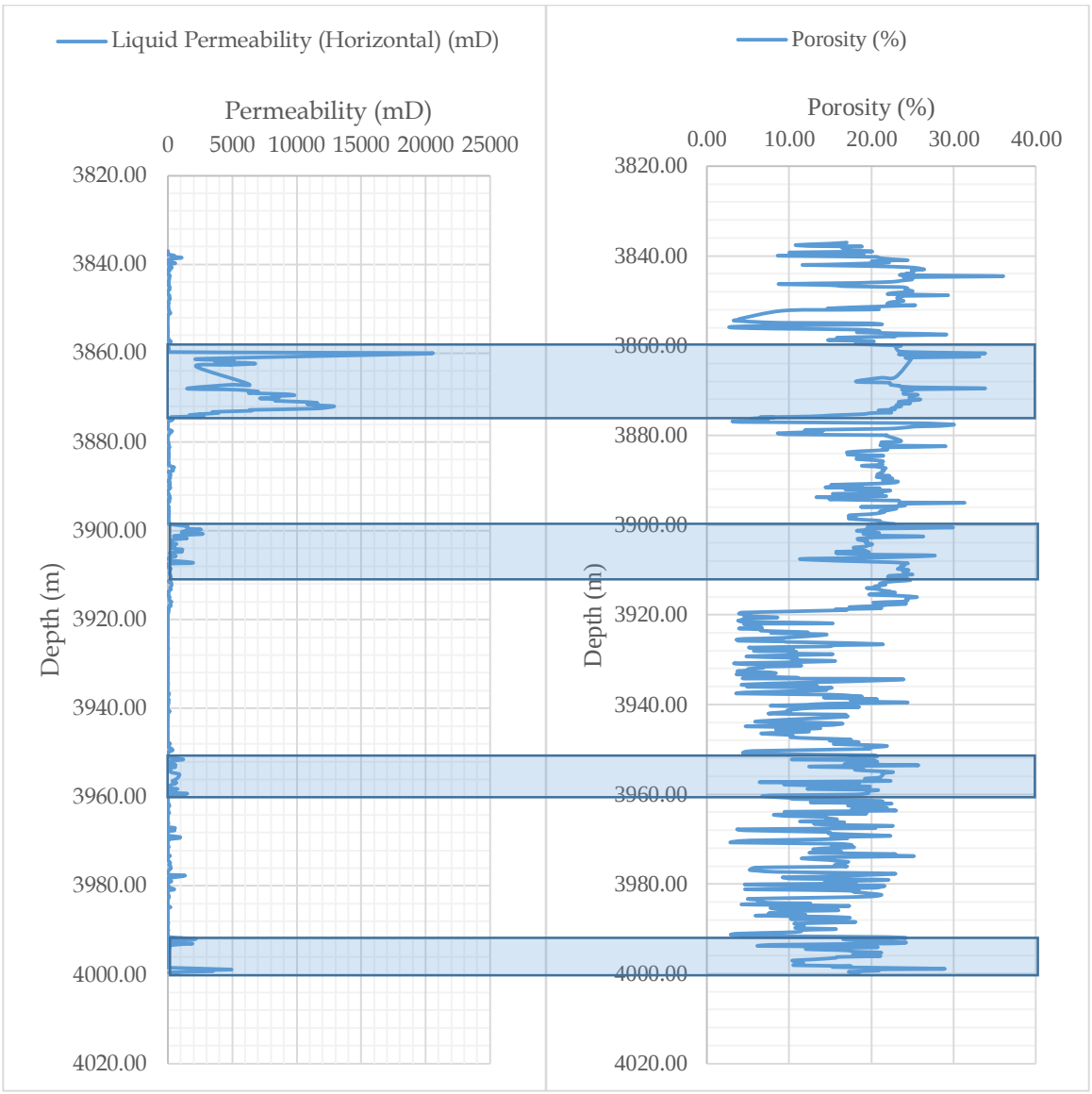


Figure 5.2 Possible CO_2 Sequestration Intervals.

5.2. Future Work

More research would be needed to fully characterize a CO_2 sequestration site by locating a seal and a trap.

Locating the Reservoir:	**Locating the Seal**:	**Locating the Trap**:
•Good Quality Lithologies (log + core); •Favorable Permeability (log + core); •Reservoir Extension (log + seismic); •Net-to-Gross Ratio (log + core).	•Typical Seal Type Lithologies (log + core); •Low Permeability Layers (log + core); •Seal Extension over Reservoir (log + seismic); •Sufficient Thickness (log + seismic).	•Presence of Structures (seismic); •Presence of Stratigraphic Traps (seismic); •Presence of Four-Way Dip Closures (seismic); •Presence of Three-Way Dip Closures (seismic).

REFERENCES

Abdulraheem, A., Sabakhi, E., Ahmed, M., Vantala, A., Raharja, I., & Korvin, G. (2007). Estimation of Permeability from Wireline Logs in a Middle Eastern Carbonate Reservoir using Fuzzy Logic. *15th SPE Middle East Oil & Gas Show and Conference* (pp. 1-11). Bahrain International Exhibition Centre, Kingdom of Bahrain: Society of Petroleum Engineers.

Ahn, S., Park, C., Kim, J., & Kang, J. M. (2018). Data-Driven Inverse Modeling with a Pre-Trained Neural Network at Heterogeneous Channel Reservoirs. *Journal of Petroleum Science and Engineering*. doi:10.1016/j.petrol.2018.06.084

Al-Anazi, A., & Gates, I. D. (2010). On the Capability of Support Vector Machines to Classify Lithology from Well Logs. *Natural Resources Research*, 125-139.

Altintas, I. (2018). Introduction to Big Data. *The Hadoop Ecosystem: Welcome to the zoo!* San Diego, California, United States of America: Coursera.

Aminzadeh, F. (1994). Applications of Fuzzy Experts Systems in Integrated Oil Exploration. *Computers Elect. Engng*, *20*(2), 89-97.

Aminzadeh, F., & Brouwer, F. (2006). Integrating Neural Networks and Fuzzy Logic for Improved Reservoir Property Prediction and Prospect Ranking. *SEG/New Orleans 2006 Annual Meeting* (pp. 1752-1756). New Orleans: Society of Exploration Geophysicists.

An, P., Yang, X., & Zhang, M. (2018). Porosity prediction and application with multiwell-logging curves based on deep neural network. *SEG Technical Program Expanded Abstracts*, 819-823.

Anifowose, F. A., Abdulraheem, A., Al-Shuhail, A., & Schmitt, D. P. (2013). Improved Permeability Prediction from Seismic and Log Data using Artificial Intelligence Techniques. *SPE Middle East Oil and Gas Show and Conference* (pp. 1-7). Manama, Bahrain: Society of Petroleum Engineers.

Baaziz, A., & Quoniam, L. (2014). How to use Big Data technologies to optimize operations in Upstream Petroleum Industry. *21st World Petroleum Congress* (pp. 1-9). Moscow: 21st World Petroleum Congress.

Bachu, S. (2000). Sequestration of CO2 in Geological Media: Criteria and Approach for Site Selection in Response to Climate Change. *Energy Conversion & Management*, 953-970.

Bachu, S. (2002). Sequestration of CO2 in Geological Media in Response to Climate Change: Road Map for Site Selection using the

Transform of the Geological Space into the CO2 Phase Space. *Energy Conversion & Management*, 87-102.

Bagheripour, P. (2014). Committee neural network model for rock permeability prediction. *Journal of Applied Geophysics*, 142-148.

Baines, S. J., & Worden, R. H. (2004). Geological storage of carbon dioxide. *Geological Society, London, Special Publications*, 1-6.

Basri, I., Aswad, S., Suryana, & Priatama, I. N. (2018). Porosity Prediction using Multiattribute Transforms and Probabilistic Neural Networks Analysis from Limestone Formation in BSJ Gas Field, Central Sulawesi, Indonesia. *Indonesian Petroleum Association, Forty-Second Annual Convention & Exhibition*. Indonesian Petroleum Association.

Benaouda, D., Wadge, G., Whitmarsh, R. B., Rothwell, R. G., & MacLeod, C. (1999). Inferring the lithology of borehole rocks by applying neural network classifiers to downhole logs: an example from the Ocean Drilling Program. *Geophysical Journal International*, 477-491.

Benson, S. M., & Franklin M. Orr, J. (2008). Carbon Dioxide Capture and Storage. *MRS Bulletin*, 303-305. doi:10.1557/mrs2008.63

Bhatt, A., & Helle, H. B. (2002). Determination of facies from well logs using modular neural networks. *Petroleum Geoscience*, 217-228.

Busch, J. M., Fortney, W. G., & Berry, L. N. (1987). Determination of Lithology From Well Logs by Statistical Analysis. *SPE Formation Evaluation.* Society of Petroleum Engineers.

Chadwick, R., Zweigel, P., Gregersen, U., Kirby, G., Holloway, S., & Johannessen, P. (2004). Geological Reservoir Characterization of a CO2 Storage Site: The Utsira Sand, Sleipner, Northern North Sea. *Energy*, 1371-1381.

Cuddy, S. J. (1997). The Applications of the Mathematics of Fuzzy Logic to Petrophysics. *SPWLA 38th Annual Logging Symposium* (pp. 1-14). Houston, Texas: Society of Petrophysicists and Well Log Analysts.

Cuddy, S. J. (2000). Litho-Facies and Permeability Prediction from Electrical Logs using Fuzzy Logic. *SPE Reservoir Eval. & Eng., 3*(4), 319-324.

Deng, C., Pan, H., Fang, S., Konaté, A. A., & Qin, R. (2017). Support vector machine as an alternative method for lithology classification of crystalline rocks. *Journal of Geophysics and Engineering*, 341-349.

Doke, J. (2016, September 1). *MathWorks®*. Retrieved from File Exchange: https://www.mathworks.com/mat

labcentral/fileexchange/7173-grabit

Duan, Y., Li, Y., Li, G., & Sun, Q. (2016). A New Neural Network Model for Rock Porosity Prediction. *2016 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI).* Beijing, China: IEEE.

Elkatatny, S., Mahmoud, M., Tariq, Z., & Abdulraheem, A. (2018). New insights into the prediction of heterogeneous carbonate reservoir permeability from well logs using artificial intelligence network. *Neural Computing and Application*, 2673-2683.

Elkatatny, S., Tariq, Z., Mahmoud, M., & Abdulraheem, A. (2018). New insights into porosity determination using artificial intelligence techniques for carbonate reservoirs. *Petroleum*, 408-418.

Equinor. (2018, May 21). *Volve data village*. Retrieved July 26, 2018, from Equinor: https://data-equinor-com.azurewebsites.net/dataset/volve

Esmaili, S., & Mohaghegh, S. D. (2016). Full Field Reservoir Modeling of Shale Assets using Advanced Data-Driven Analytics. *Geoscience Frontiers*, 11-20.

Fang, Y., Baojun, B., Dazhen, T., Dunn-Norman, S., & Wronkiewicz, D. (2010). Characteristics of CO2 Sequestration in Saline Aquifers. *Petroleum Science*, 83-92.

Firican, G. (2018). *The 10 Vs of Big Data*. Retrieved from TDWI: https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Gale, J. (2004). Geological storage of CO2: What do we know, where are the gaps and what more needs to be done? *Energy*, 1329-1338.

Ghosh, S., Chatterjee, R., & Shanker, P. (2016). Estimation of ash, moisture content and detection of coal lithofacies from well logs using regression and artificial neural network modelling. *Fuel*, 279-287.

Gogoi, T., & Chatterjee, R. (2019). Estimation of petrophysical parameters using seismic inversion and neural network modeling in Upper Assam basin, India. *Geoscience Frontiers*, 1113-1124.

Gu, Y., Bao, Z., Lin, Y., Qin, Z., Lu, J., & Wang, H. (2017). The porosity and permeability prediction methods for carbonate reservoirs with extremely limited logging data: Stepwise regression vs. N-way analysis of variance. *Journal of Natural Gas Science and Engineering*, 99-119.

Hardisty, P. E., Sivapalan, M., & Brooks, P. (2011). The Environmental Economic Sustainability of Carbon Capture and Storage. *International Journal of Environmental*

Research and Public Health, 1460-1477.

Haris, A., Fakhri, M. K., Isniarny, N., & Riyanto, A. (2017). A Case Study of Hydrocarbon Prospect Evaluation in North Exito Field by using Seismic Attribute and Seismic Inversion. *American Institute of Physics*, (pp. 1-4).

Harrison, B., & Kennedy, M. (2002). Improving Prospect Evaluation by Integrating Petrophysical Models into the Workflow. *SCA*, (pp. 1-12).

Hassanzadeh, H., Pooladi-Darvish, M., & Keith, D. W. (2007). Scaling Behavior of Convective Mixing, with Application to Geological Storage of CO2. *AIChe Journal*, 1121-1131.

HelpSaveNature. (2017, December 18). *Role of Carbon Sequestration and its Pros and Cons You Never Knew*. Retrieved September 27, 2018, from HelpSaveNature: https://helpsavenature.com/explanation-of-carbon-sequestration-with-pros-cons

Hui, Y., Baihong, W., Yan, Z., Guangya, Z., Zhizhou, L., Fengcheng, W., . . . Qinghui, H. (2009). Distribution of Hydrocarbon Traps in Volcanic Rocks and Optimization for Selecting Exploration Prospects and Targets in Junggar Basin: Case Study in Ludong-Wucaiwan Area, NW China. *Petroleum Exploration and Development*, 419-427.

Imamverdiyev, Y., & Sukhostat, L. (2019). Lithological facies classification using deep convolutional neural network. *Journal of Petroleum Science and Engineering*, 216-228.

Intergovernmental Panel on Climate Change. (2005). *Special Report on Carbon Dioxide Capture and Storage.* Cambridge, UK: Cambridge University Press.

InternetLiveStats. (2018). *Google Search Statistics*. Retrieved July 24, 2018, from InternetLiveStats: http://www.internetlivestats.com/google-search-statistics/

Iturrarán-Viveros, U., & Parra, J. O. (2014). Artificial Neural Networks applied to estimate permeability, porosity and intrinsic attenuation using seismic attributes and well-log data. *Journal of Applied Geophysics*, 45-54.

Jamshidian, M., Hadian, M., Zadeh, M. M., Kazempoor, Z., Bazargan, P., & Salehi, H. (2015). Prediction of free flowing porosity and permeability based on conventional well logging data using artificial neural networks optimized by Imperialist competitive algorithm - A case study in the South Pars gas field. *Journal of Natural Gas Science and Engineering*, 89-98.

Jamshidian, M., Hadian, M., Zadeh, M. M., Kazempoor, Z., Bazargan, P., & Salehi, H. (2015). Prediction of free flowing porosity and

permeability based on conventional well logging data using artificial neural networks optimized by Imperialist competitive algorithm – A case study in the South Pars gas field. *Journal of Natural Gas Science and Engineering*, 89-98.

Kapur, L., Lake, L. W., Sepehrnoori, K., Herrick, D. C., & Kalkomey, C. T. (1998). Facies Prediction From Core And Log Data Using Artificial Neural Network Technology. *SPWLA 39th Annual Logging Symposium, 26-28 May*. Keystone, Colorado: Society of Petrophysicists and Well-Log Analysts.

Kaszuba, J. P., Janecky, D. R., & Snow, M. G. (2003). Carbon dioxide reaction processes in a model brine aquifer at 200 °C and 200 bars: implications for geologic sequestration of carbon. *Applied Geochemistry*, 1065-1080.

Klinkenberg, L. J. (1941). The Permeability of Porous Media to Liquids and Gases. *Production Practice*, 200-213.

Konaté, A. A., Pan, H., Khan, N., & Yang, J. H. (2015). Generalized regression and feed-forward back propagation neural networks in modelling porosity from geophysical well logs. *Journal of Petroleum Exploration and Production Technology*, 157-166.

Konaté, A. A., Pan, H., Khan, N., & Ziggah, Y. Y. (2015). Prediction of porosity in crystalline rocks using artificial neural networks: An example from the Chinese Continental Scientific Drilling Main hole. *Studia Geophysica et Geodaetica*, 113-136.

Kovscek, A. R. (2002). Screening Criteria for CO2 Storage in Oil Reservoirs. *Petroleum Science and Technology*, 841-866.

Lewis, H. G., & Brown, M. (2001). A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, 3223-3235.

Lim, J.-S., & Kim, J. (2004). Reservoir Porosity and Permeability Estimation from Well Logs using Fuzzy Logic and Neural Networks. *SPE Asia Pacific Oil and Gas Conference and Exhibition* (pp. 1-9). Perth, Australia: Society of Petroleum Engineers.

Ma, Y. Z. (2019). Facies and Lithofacies Classifications from Well Logs. In Y. Z. Ma, *Quantitative Geosciences: Data Analytics, Geostatistics, Reservoir Characterization and Modeling* (pp. 231-254). Springer, Cham.

Mahmood, M. F., & Ahmad, Z. (2017). Application of Multi-Layer Feed Forward Neural Network (MLFNN) for the Prediction of Porosity: A Case Study from Lower Indus Basin, Pakistan. *The Nucleus*, 10-15.

Mahmood, M. F., Shakir, U., Abuzar, M. K., Khan, M. A., & Khattak, N.

(2017). Probabilistic neural network approach for porosity prediction in Balkassar area: a case study. *Journal of Himalayan Earth Sciences*, 111-120.

Mahmoodi, O., Smith, R. S., & Tinkham, D. K. (2016). Supervised classification of down-hole physical properties measurements using neural network to predict the lithology. *Journal of Applied Geophysics*, 17-26.

Malvić, T., Velić, J., Horváth, J., & Cvetković, M. (2010). Neural Networks in Petroleum Geology as Interpretation Tools. *Central European Geology*, 97-115.

Marr, B. (2015). *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance.* Cornwall: John Wiley & Sons Ltd.

Maurya, S. P., & Singh, K. H. (2019). Predicting Porosity by Multivariate Regression and Probabilistic Neural Network using Model-based and Coloured Inversion as External Attributes: A Quantitative Comparison. *Journal of the Geological Society of India*, 207-212.

MongoDB, Inc. (2018). *What is MongoDB*. Retrieved from MongoDB: https://www.mongodb.com/what-is-mongodb

Montgomery, D. C., & Runger, G. C. (2014). *Applied Statistics and Probability for Engineers* (6th ed.). Singapore: John Wiley & Sons Singapore Pte. Ltd.

Moon, S., Lee, G. H., Kim, H., Choi, Y., & Kim, H.-J. (2016). Collocated cokriging and neural-network multi-attribute transform in the prediction of effective porosity: A comparative case study for the Second Wall Creek Sand of the Teapot Dome field, Wyoming, USA. *Journal of Applied Geophysics*, 69-83.

Nashawi, I., & Malallah, A. (2010). Permeability Prediction from Wireline Logs using Fuzzy Logic and Discriminant Analysis. *SPE Asia Pacific Oil & Gas Conference and Exhibition* (pp. 1-12). Brisbane, Queensland, Australia: Society of Petroleum Engineers.

Norwegian Petroleum Directorate. (2018). *Factpages*. Retrieved from Norwegian Petroleum Directorate: http://factpages.npd.no/factpages/

Ojha, M., & Maiti, S. (2016). Sediment classification using neural networks: An example from the site-U1344A of IODP Expedition 323 in the Bering Sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, 202-213.

Olatunji, S., Selamat, A., & Abdulazeez, A. (2015). Harnessing the Power of Type-2 Fuzzy Logic System in the Prediction of Reservoir Properties. *SPE Saudi Arabia Section Technical Symposium*

and Exhibition (pp. 1-24). Al-Khobar, Saudi Arabia: Society of Petroleum Engineers.

Otis, R. M., & Schneidermann, N. (1997). A Process for Evaluating Exploration Prospects. *AAPG Bulletin*, 1087-1109.

Oyeyemi, K. D., Olowokere, M. T., & Aizebeokhai, A. P. (2018). Hydrocarbon Resource Evaluation using Combined Petrophysical Analysis and Seismically Derived Reservoir Characterization, Offshore Niger Delta. *J Petrol Explor Prod Technol*, 99-115.

Patrizio, A. (2018, May 30). *Big Data vs. Artificial Intelligence*. Retrieved from Datamation: https://www.datamation.com/big-data/big-data-vs.-artificial-intelligence.html

Qmee. (2013). *Online in 60 Seconds*. Retrieved July 24, 2018, from Qmee: https://blog.qmee.com/wp-content/uploads/2013/07/Qmee-Online-In-60-Seconds2.png

Raef, A., Meek, T. N., & Totten, M. W. (2016). Applications of 3D Seismic Attribute Analysis in Hydrocarbon Prospect Identification and Evaluation: Verification and Validation based on Fluvial Paleochannel Cross-Sectional Geometry and Sinuosity, Ness Country, Kansas, USA. *Marine and Petroleum Geology*.

Rafik, B., & Kamel, B. (2017). Prediction of permeability and porosity from well log data using the nonparametric regression with multivariate analysis and neural network, Hassi R'Mel Field, Algeria. *Egyptian Journal of Petroleum*, 763-778.

Rafik, B., & Kamel, B. (2017). Prediction of permeability and porosity from well log data using the nonparametric regression with multivariate analysis and neural network, Hassi R'Mel Field, Algeria. *Egyptian Journal of Petroleum*, 763-778.

Reservoir Rock Properties Laboratory Manual. (2012). Dhahran: King Fahd University of Petroleum & Minerals.

Roisenberg, M., Schoeninger, C., & Silva, R. R. (2009). A Hybrid Fuzzy-Probabilistic System for Risk Analysis in Petroleum Exploration Prospects. *Expert Systems with Applications*, 6282-6294.

Saggaf, M. M. (2002). Estimation of Reservoir Quality by Attribute Integration through Fuzzy Logic. *17th World Petroleum Congress* (pp. 231-241). Rio de Janeiro, Brazil: Society of Petroleum Engineers.

Saggaf, M. M., & Nebrija, E. L. (2000). Estimation of Lithologies and Depositional Facies from Wire-Line Logs. *AAPG Bulletin*, 1633-1646.

Saljooghi, B. S., & Hezarkhani, A. (2015). A new approach to improve permeability prediction of petroleum reservoirs using neural network adaptive wavelet (wavenet). *Journal of Petroleum Science and Engineering*, 851-861.

Sammut, C., & Webb, G. I. (2017). *Encyclopedia of Machine Learning and Data Mining.* New York: Springer Science. doi:10.1007/978-1-4899-7687-1

Silva, A. A., Neto, I. A., Misságia, R. M., Ceia, M. A., Carrasquilla, A. G., & Archilha, N. L. (2015). Artificial neural networks to support petrographic classification of carbonate-siliciclastic rocks using well logs and textural information. *Journal of Applied Geophysics*, 118-125.

Singh, S., Kanli, A. I., & Sevgen, S. (2016). A general approach for porosity estimation using artificial neural network method: a case study from Kansas gas field. *Studia Geophysica et Geodaetica*, 130-140.

Son, S., Hou, J., Liu, Y., Cao, S., Hu, C., Wang, X., & Chang, Z. (2016). Application of artificial neural network in Geology: Porosity estimation and lithological facies classification. *12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. Changsha, China: IEEE.

Spofforth, D., & Firth, J. (2016). Integrated Datasets Hold the Key to Unravelling Petroleum Prospectivity. *World Oil® Special Focus: Advances in Exploration*, 33-38.

Srinivasan, S., & Petkovic, D. (2000). Phonetic confusion matrix based spoken document retrieval. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 81-87). Athens: ACM Digital Library.

Taghavi, A. (2005). Improved Permeability Estimation through use of Fuzzy Logic in a Carbonate Reservoir from Southwest Iran. *14th SPE Middle East Oil & Gas Show and Conference* (pp. 1-9). Bahrain International Exhibition Centre, Bahrain: Society of Petroleum Engineers.

Tang, H., & White, C. D. (2008). Multivariate statistical log log-facies classification on a shallow marine reservoir. *Journal of Petroleum Science and Engineering*, 88-93.

The Apache Software Foundation. (2008). *HDFS Architecture Guide*. Retrieved from Hadoop: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

The Apache Software Foundation. (2013). *MapReduce Tutorial*. Retrieved from Hadoop:

https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

The Apache Software Foundation. (2014). *Apache Hive*. Retrieved from Apache Hive: http://hive.apache.org/

The Apache Software Foundation. (2015). *Apache Storm*. Retrieved from Apache Storm: http://storm.apache.org/index.html

The Apache Software Foundation. (2016). *What is Cassandra*. Retrieved from Apache Cassandra: http://cassandra.apache.org/

The Apache Software Foundation. (2017). *Welcome to Apache Zookeeper™*. Retrieved from Apache ZooKeeper™: http://zookeeper.apache.org/

The Apache Software Foundation. (2017). *What is Apache Flink?* Retrieved from Apache Flink: https://flink.apache.org/flink-architecture.html

The Apache Software Foundation. (2018). *Apache Hadoop YARN*. Retrieved from Apache Hadoop: https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

The Apache Software Foundation. (2018). *Welcome to Apache Giraph!* Retrieved from Apache Giraph: http://giraph.apache.org/

The Apache Software Foundation. (2018). *Welcome to Apache HBase™*. Retrieved from Apache HBase: http://hbase.apache.org/

The Apache Software Foundation. (2018). *Welcome to Apache Pig!* Retrieved from Apache Hadoop: http://pig.apache.org/

The Apache Software Foundation. (2018). *Welcome to Apache™ Hadoop®!* Retrieved from Apache Hadoop: http://hadoop.apache.org/

The MathWorks, Inc. (2009). *Neural Network Toolbox™ 6 User's Guide.* The MathWorks, Inc.

The MathWorks, Inc. (2010). *Fuzzy Logic Toolbox™ 2 User's Guide.* The MathWorks, Inc.

Torp, T. A., & Gale, J. (2004). Demonstrating storage of CO2 in geological reservoirs: The Sleipner and SACS projects. *Energy*, 1361-1369.

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, 40-50.

U.S. Department of Energy. (2016). *Carbon Capture, Utilization, and Storage: Climate Change, Economic Competitiveness, and Energy Security.* U.S. Department of Energy.

University of Birmingham. (2018). *How to Write a Research Proposal*. Retrieved from University of Birmingham: https://www.birmingham.ac.uk/schools/law/courses/research/research-proposal.aspx

Vidas, H., Hugman, B., Chikkatur, A., & Venkatesh, B. (2012). *Analysis of the Costs and Benefits of CO2 Sequestration on the U.S. Outer Continental Shelf*. U.S. Department of the Interior, Bureau of Ocean Energy Management.

Visa, S., Ramsay, B., Ralescu, A., & Knaap, E. v. (2011). Confusion Matrix-based Feature Selection. *Proceedings of the Twenty--second Midwest Artificial Intelligence and Cognitive Science Conference* (pp. 120-127). Cincinnati: Omnipress - Madison, WISCONSIN.

Walls, J. D., Taner, M. T., Guidish, T., Taylor, G., Dumas, D., & Derzhi, N. (1999). North Sea reservoir characterization using rock physics, seismic attributes, and neural networks; a case history. *SEG Technical Program Expanded Abstracts* (pp. 1572-1575). Society of Exploration Geophysicists.

West, J. M., Pearce, J., Bentham, M., & Maul, P. (2005). Issue Profile: Environmental Issues and the Geological Storage of CO2. *European Environment*, 250-259.

What is an ROC curve? (n.d.). Retrieved from The Analysis Factor: https://www.theanalysisfactor.com/what-is-an-roc-curve/

Workflows for Data Science Center of Excellence. (2018). *Big Data*. Retrieved July 24, 2018, from Workflows for Data Science Center of Excellence: https://words.sdsc.edu/words-data-science/big-data

Yelebe, Z. R., & Samuel, R. J. (2015). Benefits and Challenges of Implementing Carbon Capture and Sequestration Technology in Nigeria. *The International Journal of Engineering and Science*, 42-49.

APPENDIX

Table A.1 MATLAB code used to conduct statistical analysis on input data.

```
%Statistical Analysis for Well 15_19-19 A
GR = xlsread('Well15_9_19A_Analysis.xlsx','Training-Testing Dataset','W3:W559');
RHOB = xlsread('Well15_9_19A_Analysis.xlsx','Training-Testing Dataset','X3:X559');
NPHI = xlsread('Well15_9_19A_Analysis.xlsx','Training-Testing Dataset','Y3:Y559');
PERM = xlsread('Well15_9_19A_Analysis.xlsx','Training-Testing Dataset','Z3:Z559');

%Statistical Analysis for GR
GR_Minimum = min(GR);
GR_Maximum = max(GR);
GR_ArithmeticMean = mean(GR);
GR_GeometricMean = geomean(GR);
GR_HarmonicMean = harmmean(GR);
GR_Mode = mode(GR);
GR_Range = range(GR);
GR_MidRange = (GR_Maximum + GR_Minimum)/2;
GR_Median = median(GR);
GR_Variation = var(GR);
GR_IQR = iqr(GR);
GR_StdDev = std(GR);
GR_Skewness = skewness(GR);
GR_Kurtosis = kurtosis(GR);
GR_FrequencyHistogram = histogram(GR);
GR_Covariance = cov(GR);
GR_CoeffVariation = GR_StdDev/GR_ArithmeticMean;
GR_CorrCoeff = corr2(GR,PERM);
GR_CorrCoeff2 = corr2(GR,log10(PERM));
GR_CorrCoeff3 = corr2(log10(GR),log10(PERM));

%Statistical Analysis for RHOB
RHOB_Minimum = min(RHOB);
RHOB_Maximum = max(RHOB);
RHOB_ArithmeticMean = mean(RHOB);
RHOB_GeometricMean = geomean(RHOB);
RHOB_HarmonicMean = harmmean(RHOB);
RHOB_Mode = mode(RHOB);
RHOB_Range = range(RHOB);
RHOB_MidRange = (RHOB_Maximum + RHOB_Minimum)/2;
RHOB_Median = median(RHOB);
RHOB_Variation = var(RHOB);
RHOB_IQR = iqr(RHOB);
RHOB_StdDev = std(RHOB);
RHOB_Skewness = skewness(RHOB);
RHOB_Kurtosis = kurtosis(RHOB);
RHOB_FrequencyHistogram = histogram(RHOB);
RHOB_Covariance = cov(RHOB);
RHOB_CoeffVariation = RHOB_StdDev/RHOB_ArithmeticMean;
RHOB_CorrCoeff = corr2(RHOB,PERM);
RHOB_CorrCoeff2 = corr2(RHOB,log10(PERM));
RHOB_CorrCoeff3 = corr2(log10(RHOB),log10(PERM));
```

```
%Statistical Analysis for NPHI
NPHI_Minimum = min(NPHI);
NPHI_Maximum = max(NPHI);
NPHI_ArithmeticMean = mean(NPHI);
NPHI_GeometricMean = geomean(NPHI);
NPHI_HarmonicMean = harmmean(NPHI);
NPHI_Mode = mode(NPHI);
NPHI_Range = range(NPHI);
NPHI_MidRange = (NPHI_Maximum + NPHI_Minimum)/2;
NPHI_Median = median(NPHI);
NPHI_Variation = var(NPHI);
NPHI_IQR = iqr(NPHI);
NPHI_StdDev = std(NPHI);
NPHI_Skewness = skewness(NPHI);
NPHI_Kurtosis = kurtosis(NPHI);
NPHI_FrequencyHistogram = histogram(NPHI);
NPHI_Covariance = cov(NPHI);
NPHI_CoeffVariation = NPHI_StdDev/NPHI_ArithmeticMean;
NPHI_CorrCoeff = corr2(NPHI,PERM);
NPHI_CorrCoeff2 = corr2(NPHI,log10(PERM));
NPHI_CorrCoeff3 = corr2(log10(NPHI),log10(PERM));
```

Table A.2 MATLAB code used to extract weights and biases from developed ANN model.

```
save net;
weights_biases = getwb(net);
[b1,IW,LW] = separatewb(net,weights_biases);
```

VITAE

Name : Muhammad Said Abdallah

Nationality : Tanzanian

Date of Birth : 13th of February 1995

Email : muhammad.said512@gmail.com

Address : P.O. Box 742, KFUPM, Dhahran 31261, Saudi Arabia

Academic Background : MSc. in Environmental Sciences

BSc. in Petroleum Engineering

Publications & Presentations :

Sauerer, B., Abdulghani, A. J., **Abdallah, M. S.**, & Abdallah, W. (2018). Interfacial Activity of Characterized Middle Eastern Asphaltenes. Energy & Fuels. doi:10.1021/acs.energyfuels.8b03085

Bastian Sauerer, Abdullah J. Al Abdulghani, **Muhammad S. Abdallah**, Wael Abdallah, The 19th Annual Conference on Petroleum Phase Behavior and Fouling, The Chateaux, Deer Valley, UT, USA, July 8-12, 2018. Technical Presentation: "Interfacial Activity of Characterized Middle Eastern Asphaltenes".

Abdallah, M. S., Al-Zawad, M. A., & Fraim, M. L. (2017, June 1). Common Misinterpretations of Gas Condensate Reservoirs. Society of Petroleum Engineers. doi:10.2118/187995-MS